Toward a Context-Driven Model of WWW Navigation

Final Report

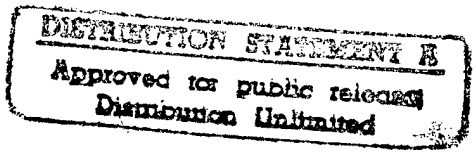
October 2, 1997

Office of Naval Research
Contract #: N00014-97-C-0108

ROI Joint Venture
15510 Seahorse, Houston TX 77062

Dr. Susan L. Gerhart, Managing Partner
281-486-8480 slger@netropolis.net

Ted Ralston, Business Director
206-820-6663 tedr@mindspring.com



"Toward a Context-Driven Model of Web Navigation

Summary of Results

This research addresses problems of current WWW users by experimenting with alternative techniques for accessing and analyzing web information. The objective is to increase the productivity of web users and to improve their ability to find, qualify, and propagate high quality materials. The key idea of "Browsing in Context" is: *user-managed alternative views of collections of web materials provide higher-level insights on trends and patterns within the collections and improved direct interaction with abridged and full materials.*

"Browsing in Context" (BiC) offers a different approach to accessing and using WWW information content. Current browsers, commercial web utilities, and desktop computing systems do not adequately support several needs of WWW Information Professionals (librarians, journalists, market and policy analysts, research program trackers, etc.). Specifically lacking are information management and analysis tools for Topic Search Management, i.e. collecting and organizing large collections of URLs on specific subjects.

• Our specific results are as follows:

1. Refined definitions of 3 views of URL collections that support the "browsing in context" model, where the user establishes known regions of the URL collection according to various perspectives (links, domains, and concepts) to browse and evaluate the materials. Appendix A contains a position paper for the CHI 97 Workshop on "Augmented Conceptual Analysis of the Web".
2. Empirical characterization of web materials on a variety of subjects. This helps establish a baseline of requirements for tools and expectations for users and relates our approach to information foraging.
3. A scenario, cast as a survey, that we believe captures the essence of requirements for web information professionals. Appendix B contains the survey on the subject of "access control and security policies for digital library data" and instructions for using HTML summary reports (available upon request).
4. A survey of research literature that informs the BIC model and provides additional attacks on the same general problems of improved web user productivity.
5. A preliminary casting of the above results in the form of "design patterns", a recent approach to organizing common computing and business procedures at an abstract level, stemming from the object-oriented modelling techniques.

The underlying technology used for the experimentation is LodeStar, described in Appendix C with an accompanying screenshot demonstration on the website. Appendix D contains an example, "information foraging", output from the LodeStar toolset. The reference list consists of LodeStar-produced index files of URLs by Internet Domain and URLs by Report File.

Further work could include: cost-benefit and formal architectural models for alternative modes of browsing in context, establishing a baseline topic area and community collection mechanism for a specific topic area, augmented analyses by document comparison and other information retrieval techniques.

Where Current Tools Fail the Web Information Professional

The WWW presents two sets of problems. First is sheer size and growth, covering both quantity of materials and frequent appearance and disappearance of materials. Second, is the lack of metadata and organization that permits classification by type (technical, commercial, promotional, personal, etc.)

Today's technology -- browsers, search engines, bookmark managers, and word processors -- fail at the point where web users need:

- alternative views of their collections to better understand what they have, how the materials are distributed, what might be missing, and what might be most valuable and
- editing operations and generating reports for themselves and other users.

Search engines (Lycos, Altavista, etc.) emerged to cope with size issues and have begun to tackle the classification issue, both by using human classifiers and evaluators (Magellan, Yahoo) and algorithmic clustering (Altavista, NLSearch). The problems of using search engines are well documented (e.g. columns and tutorials). Crafted URL lists provided by subject experts or aficionados provide an alternative by winnowing down larger topics.

Browsers that began as information viewers have evolved little in that direction, remaining weak in their history, bookmark management, and organizational mechanisms. They deliver HTML files to the desktop and provide rendering of the HTML, leaving the organization of materials up to the web user. A marketplace of web utilities have emerged to assist the web user in using search engines and in storing and processing materials. Metasearchers such as EchoSearch and WebFerret query

multiple search engines, merge results, and provide selectable lists and automatic downloading.

Thus, it is possible to, in minutes, amass a considerable amount of web pages on a selected topic using automated downloaders from search engines, website crawlers, and selected collections, thereby generating new problems in filtering, organizing, and propagating collection results. Given that there are several hundred URLs on a particular subject, what techniques and tools help the WWW IP to understand and process the URLs toward their goals at the moment?

Currently, there are few analytic tools available for lists of URLs and only rudimentary editing tools. Altavista and Northern Lights both provide elementary clustering of URLs by type of site, by discipline (e.g. biology), person, place. GUI mechanisms such as Windows 95 ListView offer multiple-columns so that URLs may be listed then sorted by date, URL, title, etc. and deleted.

To address this gap, we are using as an underlying toolset for the experiments performed under this contract the LodeStar tools described in Appendix C. These tools offer the net access, HTML rendering, bookkeeping, visualization, and rudimentary analysis to support browsing in context from the perspectives of: Internet Domain, Links, and Concepts. However, these tools have not yet been user-tested.

Compounding the bookkeeping and other information management problems associated with the scale of the Web is the lack of metadata to assist classification and organized access to the Web. We do not directly address this issue but expect to see slow improvement through the work of Digital Library Metadata groups (see summaries in the Digital Library Newsletter). A significant realization that appears clearly in our empirical studies is that the Web has some aspects of a traditional technical library but perform especially effectively at delivering the "gray literature" of product descriptions, organizational newsletters, and other traditionally unmanageable materials. Any web classification improvements would provide more contexts for our model, e.g. product descriptions and reviews.

Our claim, supported by the tool prototypes and partially validated by these experiments, is that the browser component (in the narrow sense of the HTML rendered) should be a subordinate to a user-directed information management system rather than the current market game of browsers dominating the desktop and becoming operating systems. This inverted mode of thinking gives rise to new strategies for browsing, such as described here.

Browsing In Context -- an Information Foraging Model

BiC is a foraging strategy. In this section, we clarify its appliability (tppic collections) and identify some of the "laws" of information space we've observed.

As an example of the representations used during BiC, Appendix D provides an annotated outline for the .org domain "information foraging".

Information Foraging Actions - What do Web Users Do?

Consider the following types of Web user tasks:

Just Browsing - looking through, possibly complex material, without clear goals

Searching - looking for some fact, particular type of material, or specific file or location

Collecting - identifying materials on a particular topic and saving pointers and/or files

Qualifying - assuring oneself or another that a particular item meets criteria such as topic inclusion, quality standards, or due diligence requirements

Scoping - defining the boundaries of a topic, especially with respect to a given collection

Rating - applying a set of criteria to a given item, placing it in define categories

Dividing - splitting a collection according to some qualifying, scoping, rating, or collecting requirement

Merging - brining together and identifying common elements in different collections

Slicing - focusing items of a collection on a particular aspect of the collection's subject

Reporting - providing descriptions of content and overall collection structure

Updating - identifying materials that have changed and how so, new materials on a topic, and taking action on changes

We use the term "browsing" to encompass all the aspects of the above activities that include going from one item to the next to perform needed actions. For example, browsing for updating means going from item to item (possibly skipping many) to find those items with changes requiring either notes of change or replacement with changed items.

The emphasis of "browsing in context" is that the browser (person) is either

1. building a mental model of the collection and its attributes using different views and analyses, or
2. taking action on items within some known "context", such as filtering, rating, annotating, displaying, etc.

Of course, there's always a "context" of some sort, e.g. a search engine's results list is a product of its index, query, and relevance mechanism, possibly also influenced by the load on its system, advertising pressure, etc. In our model, rather, the browser (person) is operating in contexts they have imposed upon the material and therefore have better, or further control of.

Contexts in our terms are basically orderings that provide:

continuity - one item has something definite in common with or different from its predecessors or successors

extremes - there's a place to start and a place to end and progress through items can be measured

visualizations - ways of showing grouping

The browser (person) processing large collections of materials gains from contexts the ability to focus on defined features and then process similar items together, with similarity being well-defined and kept in mind. Thus, the browser (tool) should support

the browser (person) need for context.

Information Foraging Propositions

The objective of information foraging (see Appendix D and accompanying diskette) is to discover regularities in information intensive activities by looking upon the problem in ecological terms - a system of materials supporting a predator-prey interplay of activities. The raw materials are information items distributed, created, modified, duplicated, etc. across networks. The predator-prey relationship characterizes the browser (person) seeking bits of good materials in rich or diverse patches, i.e. foraging.

Our model is basically a two-phase approach. First, the browser identifies possible desirable items, then acquires them for off-line processing. The second phase is where we place our emphasis, how the "offline", i.e. already collected and localized, activities may be supported to better define collections to maintain and use. Perhaps this is like collecting seeds, or genes, with the right parentage, features, and possibilities, while trying to understand the total pool and each item's relationship to the whole, also encouraging wild and singular items as members that broaden the scope and overall evolution of the total pool.

The Information Foraging approach is as much a research methodology as it is a metaphor. The methodology includes attempts to mathematically define the regularities as laws that offer predictability and explanation. Experiments are then performed to better understand the effects of the regularities on systems and users with intent to improve system's performance, develop better tools, and provide training to users.

We have also identified a few regularities (amplified on in empirical baselines), although these are far from adequately measured nor are they statistically characterized.

- **"Law of Dead URLs"**, *10-20% of just about any collection of URLs will be inaccessible*,

suggests that a part of each phase above (acquisition, evaluation) must be to identify and resolve action for dead URLs. In fact, while many URLs are indeed dead, many are simply mutated (reorganized server file system) or moved. In many cases, it is easy to resolve a dead URL by visiting the site and either finding the changed URL or looking it up in an archive. the significance of this action is in its cost: 100 dead URLs in a collection of 1000 x 10 minutes per URL = a day's work. Recognizing the existence of this law helps estimate cost of acquiring a collection and valuating its items (what if the missing URLs are from a key site?).

- **"Law of Off-Topic URLs"**, *10-20% of URLs delivered from search engines are products of linguistic confusion*,

is both a fact of life and a result of rapid commercial transition of information retrieval technology into widespread use without the benefit of library-style metadata. \While some search systems are extended from technology regularly evaluated for precision and recall, most of the major commercial search engines have evolved from a mixture of technology concerns and commercial opportunities. There is no particular reason a search engine company should meet higher than necessary standards when its revenue is derived from advertising and auxiliary information sources. That metadata (library-style classification, even mundane details such as authorship and date) is lacking comes from the rapid growth of the Web, particularly in the realm of "gray literature" (products/services literature, individual publications). The burden then imposed upon serious web users is (1) acquire skills and expectations to manage search engines effectively and (2) find and use tools that assist in the resolution of off-topic URLs.

There may well be tradeoffs in the process of acquiring and qualifying collections of URLs, e.g. using a metasearcher may dig deeper into search engines far more rapidly than individually entered queries, but at the cost of more off-topic URLs because it's not "one query fits all". While we do not have an accurate characterization of the range for either of these laws, we can consider the consequences of either broadening the range or extending the boundaries. If only 1% of URLs in a collection were dead, we might still have to look at all 100% to determine that 1% but we would probably move the aliveness test somewhere later in the process. If 50% of all URLs were dead, we probably would be looking for better search engines or giving up our goals. The estimates we have made suggest that the problem is in the manageable, but costly range.

- **"Law of Quality Materials"**, *about 1/2 of any collection can be eliminated without much loss*,

is more subjective but again suggestive of actions. Considering that quality is both a matter of objectives for the collection and subjective to the evaluator, the main benefit of this estimate is its suggestion that better evaluation tools are required.

This is the primary hypothesis of the experiment we designed, that a more managed process of collection and evaluation using the "browsing in context" approach will produce better results for a variety of types of standard tasks. For example, we've identified the frequent occurrence of collections with perhaps 5% of the URLs being job postings or resumes, particularly where an industrial practice or technology is involved. Since the existence and quantity of such URLs may tell us something about the vitality and currency of the subject, these URLs help characterize a topic collection but not in any

particularly usable or enduring way. Thus, a job-filtering function of an evaluation system would be helpful, if not required. Literature on Information Foraging

Finally,

- **"Law of Disconnection"**, *few (maybe 5%) of pages are truly part of a web*,

derives from the behavior of the web population and suggests ample opportunity for organizational improvement. Consider the academic population, where pages are posted for self-reference with, increasingly, pointers to directly related work available online. But that's still a small proportion of the total links that could be drawn among web pages. A real web of a subject would have 100s of links between any two pages, e.g. in common terms used. More evolved argumentation structures such as IBIS (issue-based information systems) might draw out more significant relationships among documents, e.g. issue-position-argument or refines-generatizes or theory-practice or implements-evaluates.

Thus we see a range of questions that our "browsing in context" approach and tools can address, although imprecisely and, currently, informally. Of most interest to us in the future are, at the lower-level, trade-offs and, higher, stronger intellectual relationships that derive from hypertext and argumentation.

Information Foraging Technology

We discuss this further in the literature review, but a comparison is worth discussing here, namely alternative ways of using text analysis to assist foraging.

We have not yet experimented with text-based clustering but rather have found useful an explicit and directly manipulated representation of terms from documents. While clearly limited in scale, we are finding, using the SWAPI technology, some 30,000+ concepts in a 1000 URL collection. Our Concept Browser permits us to scroll up and down on a compressed list of concepts to find ones of interest and we can also order concepts by frequency. A concept can then be "browsed" across occurrences within the documents by selecting a summarized or few surrounding words of interest and then navigating to that point in documents. This strategy is more one of bottom-up, small clusters in contrast to the algorithmic feature vector clustering of scatter-gather. An experiment might well be defined here, e.g. to determine how large a concept list can be managed with this type of direct representation and manipulation, how users identify concepts of interest, how they process items once found to be of high or low quality based upon a term usage, and how this approach could be improved with thesaurus, glossary, and other language aids.

An anecdote may illustrate how this approach has worked well. On a due-diligence consulting assignment, the online documents from a research center was downloaded and indexed.

1. The consultants needed a quick answer to the following question: "what kind of query mechanism does the evaluated technology use?" A foray online to the center and use of its online system would have taken several minutes while the downloaded and indexed collection was immediately available. It happened that the index for 'query' (quite a lot of terms) lead immediately to the page where the query language was described. It is not clear that a clustering approach would have lead as quickly to the answer as the complete displayed index of terms and it is also clear that pre-processing and prior acquisition of materials was a large factor.
2. At another point, the investment firm for whom the technology was being assessed needed a list of the participants of the research center. "Who might have licensed the technology or had additional inside information?" "How much interest had the technology already attracted from industry as potential licensors, partners, or users?" It happened that these pages were scattered around the website, but easily identified using LodeStar's VCR-like navigation of web pages (feeding URLs into a browser), its marking operation (selecting ones to display), and directory (copying all the interesting files into a subdirectory). Within a few minutes of browsing the Domain outline, the materials were on the fax to the customer.

The main issue here is cost: of downloading and storage (2 hours, 20 MB), of user time (this instance was fully automated, with the user entering the URL, then copying the file to portable disk drives), of pre-processing (again automatic but consuming some time and planning). Other values include variable use of the materials (unanticipated queries), learning (seeing the kind of content and names of participants by browsing the concept list), Scoping (seeing that the center had multiple, interrelated research projects on medical informatics), and slicing (finding all the pages that related to the center's industrial affiliate program).

We believe that foraging is only part of the overall tasks for web users, also including rather mundane bookkeeping, file management, report generation, waste reclamation, etc.

Empirical Baseline

In the following discussion, we assume that topic collections are the main objective, either as objects in themselves (e.g. for continued reference in a library), as a basis for undefined questions, as audit trails (e.g. due diligence), to understand patterns of knowledge (e.g. influences of a specific research programs), or to observe characteristics of web materials to develop better tools, techniques, and training. That is, we contrast our scope with that of the ad hoc query posted to a search engine, with the web page author, and with the general Internet technology developer.

The following data is not statistically valid, nor is it possible to validate the data since the notion of a subject's exact content accessible on the web at a given point in time is not precisely definable. Rather we attempt to define and illustrate with experimental collections guidelines.

Quantity - How much is on the web about topic X?

Given topics vaguely characterized like "Therac-25" (a software safety case study) "FMEA" (Failure Mode and Effects Analysis, an engineering technique), "design patterns" (the common procedure representation) ,"security policy" "access control list" "digital library" (the subject of our experiment), "information foraging" (a major related research area), XX (well-known computer scientist), we often want to know:

- How many URLs are known on this subject to search engines?
- How many web pages at those URLs are actually available?
- How much content (MB, words, concepts) is there?
- How many search engine results are actually on target?
- Are some URLs referred to more often than others? Which sites contribute the most URLs?
- How many URLs are worth keeping on a list to be used for serious research?

Our composite answer is:

- Roughly 1000-3000 URLs will be delivered from queries to search engines, with very little overlap among results (because search engines index different parts of the web, different parts of sites, different parts of pages, and they visit web pages at different intervals). Further inspection will likely show that one or two search engines will have indexed the subject particularly well or have retrieval and relevance mechanisms well suited to the query.
- 10-20% of the URLs will be dead or changed. Newsletters and job postings change frequently by nature and companies and organizations shifting servers is common compounded by infrequent search engine web crawling patterns.
- 30-50MB of web content is common, with a few large (1MB) files (often from government agencies, dumped databases, or online thesauri). A commercial indexing package we use, Iconovex SWAPI, displays often 30,000 meaningful phrases (person and place names and "concepts").
- Most URL lists, after removing dead links, will reduce by 1/2 when common professional criteria are applied to define subject relevance and acceptability. Such criteria include: reducing redundancy, establishing authenticity, content worth looking at again, utility of the material, intelligibility, added value to collection, etc.
- Most URLs are linked to by at most one other page, usually from their own site. Clearinghouse and general subject classifications are often widely linked , while major textbooks and influential individuals are often referred to but without linking. In other words, the web isn't much of a real web, but rather a cluster of sites linked to primarily by and through search engines.

Now consider the requirements for tools to permit a web professional to process a subject on the order of hours given this type of characterization:

1. Metasearchers to probe multiple engines and collate results. Note that it's not, unfortunately, "one query fits all" so metasearchers will introduce additional bogus hits beyond what search engines normally deliver. Alternatively, the user can query several engines directly, saving the multiple pages delivered (10 URLs per page by Altavista through 100 by Hotbot) up to search engine limits (e.g. 500 for Infoseek), but this is extremely time-consuming and subject to loss (naming pages or finding them in the browser's cache). But to adequately collect pages on a subject, multiple search engines must be used.
2. Bandwidth, disk storage and compression to manipulate files. Meeting these limits is not difficult with current technology, but nevertheless a new process (workflow) is introduced into the web IP's toolset -- keeping track of large, time-intensive collections of files, changes to them, and directories.
3. Tools for scanning such large amounts of materials, both in abbreviated form and as full documents, to determine what is on target and to ascertain quality.
4. User skills for making sufficiently rapid decisions as to relevance and quality, means of recording and executing

decisions.
5. Ways of reporting results to users next in line, e.g. library patrons, due diligence auditors.

Appendix A contains an earlier chart of data from studies conducted prior to March 1997. Studies performed since then indicate several changes:

1. A more powerful, if sometimes erratic, metasearcher, WebFerret Pro, provides more URLs than we had been getting with other metasearchers or directly querying and clicking down the pages. It's likely there are twice as many URLs as quoted in these tables, although many of the additional ones are accounted for my duplication (mirror sites on the web) and variations of URLs.
2. Several studies have shown that key people are good indicators of the size and scope of a subject. Web-active individuals will typically refer to all the conferences they attend, the papers they present, their organizational and work affiliations, and other researchers in the field. Here are approximate numbers of URLs on several individuals:
    - David Eichmann, university professor, NASA technology transfer project leader, and web researcher - 1000+ URLs, 1/2 project reports, 1/2 references and external papers
    - Greg Notess, reference librarian, writer and columnist for Online magazines, 500+ URLs, mostly self-posted, but many references to tutorial materials from other libraries
    - Anita Jones, former DoD Director of Research and Engineering, 500 URLs, mostly technology policy and public statements
    - Harlan Mills, industrial research leader, founding new approaches to software engineering, 400 URLs, on case studies and technology transfer of his approach
3. From our survey,"security policy" + "access control lists" + "digital libraries", roughly 1500 URLs, probably upward of 2000 for security-related pages.

Why is this information significant? If the rule "500-1500 per specific topic" holds, then we know

- better how much data our tools must handle, a key issue for visualization tools
- Web searchers must not stop with single search engines if they expect thorough coverage of a subject (although often their objective is a specific reference or answer)
- some growth will occur but it's likely that for many topics much of the material is already online

For the purposes of experimenting with "browsing in context", this tells us that, generally, we've validated that we're in a worthwhile range of exploration:

- there is sufficient material on most subjects to warrant multiple views.
- there is not so much material available that we cannot begin experiments with our existing tool infrastructure although we might expect to stress these or outgrow them soon.
- multiple contexts will exist and be interesting.
- the amount is beyond what normal users would attempt by hand using existing browsers, word processing, and other tools

Quality - What should go into a good topic collection?

While quantitative data is relatively easy to come by, how can we address any baseline for quality? So far, we've performed two exercises in comparative analysis of collections.

For FMEA, we worked with a graduate computer science student on her independent study project on software safety using Nancy Leveson's Safeware book. She applied her background in civil engineering to identify key high-content resources on FMEA. Working with only the reports, she expressed considerable frustration at getting to most sites on the web (even at 2 a.m.) but worked better with the downloaded files. She assessed about 10 pages as being highly valuable and interesting, including some Stanford technical reports, European Esprit project descriptions, and a few product-related discussions.

In another earlier project, we had collected and then assessed several hundred pages on Cocomo II, a software cost model, for a colleague selling a COCOMO implementation. My assessment of the value of a link to put on his website, as well as identification of competitors, agreed with his on about 85% of the URLs. That we could provide this type of assessment at all required a process and tools for capturing rating information.

Information Quality is a major discussion topic for many newsgroups, ranging from website design to quality of information. The best central sources we've found are two mailing lists: UCSD Professor Phil Agre's Red Rock Eater List and INFORMATION QUALITY WWW Virtual Library

A Survey/Experiment - Discussion

What do Web Professionals do? How do they currently work? How could they use "browsing in context"?

Description of Experiment

We designed an experiment and survey (see Appendix B) to address these questions. The experiment consists of:

1. a scenario setting capturing many of the requirements for web information in organizational settings
2. a set of tasks that we believe are typical, inducing many of the practices known to be widely used (querying search engines, selecting a relevant resource, browsing for answers, saving files, etc.). What we were after was:
    - what is the workflow? do web users have a defined process?
    - where are the bottlenecks? how do web users recover from web failures (e.g. network slowness)?
    - how do web users plan their work? what happens when the time limit is hit?
3. a collected set of URLs on the subject
4. HTML reports capturing different views of the collected materials
5. a feedback form

We announced this survey to several dozen professionals over networks and newsgroups but we unable to attract any response. Although our publicity may have been lacking, we believe the main cause was simply the overwhelming nature of the problem and questionnaire: too much time, a severe requirement for introspection, and unclear uses of the results. However, we also found this subject to be confused with "access control" in the sense of libraries and parents controlling access to undesirable online materials. On more than one occasion, we've found it very difficult to get web users to introspect or express their problems; indeed, it's almost as if they lack a vocabulary to describe their work and the technology they use daily.

What we learned - general web tasks can be defined

However, the construction of the survey and the experiment was highly instructive for us, especially because the subject itself is important and of timely interest: "security policies and administrative access control for digital libraries".

Consider the general nature of the tasks:

1. **Find a definition of XX in the material available on the web. Identify one specific technique of XX.**
2. **Name one authority in the field of XX.**
3. **Find and qualify consulting or training individuals, firms, or associations we might go to for assistance.**
4. **How would we define and administer use of XX techniques?**
5. **Why have some organizations chosen the XX approach over other modes?**
6. **Find three products that provide XX and the organizations offering products for XX.**

How does "browsing in context" help prepare this collection and how does it help users? Note that we followed the LodeStar process described in Appendix C, but did not filter the content heavily. In previous iterations, we had experimented with various queries and found the one used: "security policy" AND "access control list" to be most effective and accurate, although search engines do not fully respect either phrases or Boolean operators.

How Browsing in Context Addresses These General Tasks

We observe the following from performing this collection and answer the questions ourselves, working primarily from the HTML reports:

1. Definitions are surprisingly hard to find, especially good ones. Often the definitions of "access control lists" were very product-specific without reference to a long-established technique. Browsing by concepts is clearly the way to find definitions, but even then there are many URLs using each term different ways. Our ConceptBrowser is effective here, but it is much more difficult using the HTML reports.
2. Finding an "authority" in a field requires judgment, of course. Identifying people and organizations that are prominent is a matter of counting URLs, appearance of names in concept lists (although the tools used don't stand out and company names are often lost), and weighting the results by personal knowledge. In this collection, we found no single individual but rather a few consulting organizations as the best sources.
3. Finding training and advice involves looking through Internet Domains. In .com, we're finding companies, in .org associations, in .edu consulting faculty or university centers, helped greatly by tools that classify long lists of URLs by site and domain. Again, the issue of qualification would involve personal knowledge, but considerable alternative

sources are available on the web.

4. The defining and administering question could be addressed, partially, by the surprising number of organization security policies posted online. It's quite likely a security administrator could find peers in the IT departments of these organizations. This is an excellent example of the "gray literature" at work.
5. For rationale, we're looking for case studies, where Internet domains would likely lead us to, e.g., the Digital Library newsletter, but we didn't identify anything worthwhile.
6. Several products, mostly in the category of "firewall", were easily identifiable by browsing the .com domain or looking through the ConceptList.

Taking a different perspective, here's how each of the three Views assist answering these questions. In contrast with looking at a list of 1000 URLs compiled form several lists each with a few hundred,

Domain - tells us, with exceptions and with some difficulty internationally, what type of organization we're dealing with. This type of information is essential context and usually quite effective, at least with U.S. domains, at narrowing down search for information such as training sources or technology transfer efforts.

Concepts - identifies web pages containing a term. Although thesauri and clusters could help, this relatively unsophisticated used of raw lists of terms is surprisingly effective. In contrast with using a search engine that has indexed terms, direct use of the term list for topics this size works quite well, e.g. identifying the most frequently used terms, scanning the list for vocabulary identification, and identifying off-topic subjects.

Links - While the link mechanism in the tools we used is not entirely accurate, e.g. missing some "same" links, we find it also essential to know where in a list of 1000 URLs are the sites with the most. Although site sizes are often a factor of search engine promotion tactics, multiple references to a site carry considerable significance. In this particular collection, we did not go through the multiple link counting but we did not find very much interlinking to lead us to any central resources. That too is information: this is a relatively unorganized subject, probably form lack of market, but it's possible we're not seeing the organization.

Overall, the experiment was a success but the survey a failure. We feel that this type of topic-specific exercise combined with articulating our own process and task experience has generated a possible standard for assessing web productivity and quality questions. We plan to post this, not as a survey but more as a challenge, to the Info-Quality and related newsgroups.

While the survey and instructions are delivered in Appendix B, the 20MB of downloaded files and HTML reports are also available upon request. upon request.

"Browsing in Context" - Research Literature Review

Following are the highlights of some of the papers we reviewed related to BiC. We attempted to cross many fields and sample both methodology and results. We also introduced some new "speed-reading" techniques in the process, by downloading online conference proceedings, indexing the results, and reading the papers inside-out.

Papers from the CHI 97 Conferences

"Computational Models of Information Scent-Following in a Very Large Browsable Text Collection" (Pirolli, CHI 97) provides a useful methodology framework - knowledge, adaptation, cognitive, and biological. Our work is partly Knowledge-oriented in trying to understand how tasks are formulated and to define a common set of tasks to evaluate our approach and tools. Adaptation also characterizes the need for understanding and supporting strategies within constraints of environment, tools, and costs. "Browsing in context" is definitely a strategy, to process one by one ordered items so as to (a) avoid frequent context-switching, thus lowering overload and increasing efficiency and (b) produce more consistent and reproducible evaluations.

The model-tracer experimental methodology is applicable to the twURL toolset, where events, user focus, and actions could be logged to develop a model of types and order of major operations. For example, browsing a domain view may start with a domain the user is familiar with or perhaps one most likely to be productive or perhaps based on size or other factors. Some set of probes of the data and some type of decision process could be investigated to determine how domain preference works.

Scatter-gather is the primary clustering mechanism used by the Xerox research team is an example of a strategy for helping the user find useful clusters of large numbers of documents in pursuit of other goals. In one sense, our Domain and Link Count Views provide definite clusters with a fixed results within a given pool. Scatter-gather uses text analysis and document features to determine similarity, providing a flexible way of shifting user attention from one term or feature to another.

"Sensemaker: An Information-Exploration Interface Supporting the Contextual Evolution of a User's Interests" (Baldenodado and Winograde, CHI 97) observes how users progress through information contexts while collecting resources on a topic. Sensemaker provides explicit bundling, duplicate detection, collection expansion, and viewing operations, one of which is URL-bundling, as in our Domain View. Their experiments suggest that indeed users find different view, as in tables of unbundled and different bundlings, useful but also show how difficult it is to design such experiments (what to measure, time to perform) and get useful results. Defined and variable level expansion-reduction operations would be useful for the BiC paradigm.

"Life, Death, and Lawfulness on the Electronic Frontier" (Pitkow and Perilli, CHI 97) bases an investigation on co-citation analysis, that 2 cited articles in the same paper have something in common worth noting. Not surprisingly, they found that the clustering of a university website closely followed its mutli-project orientation with a smattering of individual and common environment clusters. The paper also addresses the important issue of life-cycle of web pages, providing formulae which predict changes and survival based on usage patterns. We have found "links-to" more useful than "links-from" in our analyses. Since we simply extract all links from a web page, it's quite common to mix self-reference and advertising in with the primary results being listed. This is especially true of search engine web results but is also a pattern for more cohesive pages as well. Links-to answers the question: how did this URL get in the pool?

Other CHI 97 Activities

"Conceptual Analysis of the Web" was a workshop of 12 diverse individuals addressing new ways of working with the web. Much of the discussion centered around phenomena - what was happening on the web? how was it different? how could such phenomena be studied? Particularly interesting were the work of

- library professor Jena Bradley, comparing print and web publishing and her recent NLM project to assess the quality of online medical information
- HCI researcher John Carroll on how the Blacksburg Electronic Village was working out, e.g. who uses the online food-ordering services? not local people, but partners ordering to deliver to their student children
- Keith Instone, on maintaining web collaboration services
- Jay Bolter, on the web as remediation, an historical and media perspective

"The Great CHI 97 Browse-Off" was an effort to compare alternative tools for finding specific items within an extensive and diverse hierarchy of subjects, without using text search. The Xerox Perspective View provided a novel way of visually organizing and manipulating the hierarchy and was most effective during the contest. Our toolset, aside from its difficulty with the bulk of the data (over 7 MB), was not particularly well matched to the tasks identified by the organizers. We;re still not sure what the import of this type of browsing is nor how improved tools for it will get anybody's work done faster or better.

Papers from the WWW6 Conferences - And an Experiment Using Concept Browsing

Here we report a different way of analyzing the literature of a field, namely by browsing its concept/index.

This is an online record of actions taken on these Proceedings downloaded several months ago then processed into the ConceptBrowser. The goal was to find papers relevant to the subject of this research project report, i.e. relevant to the BiC model. More generally, this was an experiment in using concept-based, bottom-up browsing of the entire proceedings as opposed to top-down table-of-contents selection of papers to read. All the work was performed using the offline browser and links from concepts to papers. A time limit was imposed, about 2 hours to see how many papers and narrower topics would be found. This experiment was also motivated by not having the index of the proceedings immediately available.

Starting from top:

- "Martin Abadi" - he does security stuff (another subject of current interest), what's he doing in this collection (I was in transition from thinking of CHI). Paper was on the NetBill system, not relevant.
- "ACM", several proceedings listed for reference
- "Access control", a German-based Toolkit for managing web servers, relevant to our survey experiment web-based proceedings management, off the subject but interesting example of scenario analysis to define uses and data representations, software engineering approach
- "Forward concept of the browser" lead to paper: "Usability Studies and navigational aids for the WWW". This paper discusses the kinds of navigational problems users have with existing browsers and with with a template based on a popular HTML book, e.g. location of next-home-previous buttons at the top or bottom of a page, the usefulness of "home", and confusion about current browser histories (stack vs. temporal). One message for BiC is how its HTML reports are organized, e.g. what kinds of buttons are needed.
- "Altavista" lead to the WWW6Best Paper on "Syntactic Clustering of the Web", discussing the algorithms and issues behind the clustering mechanism available at the Altavista search engine. As well as the possibility of refining queries, their algorithms provide for "lost and found", distributed updating, and intellectual property protection. We see some of the same potential for our explicit concept listing approach using the search engines for retrieval then clustering and comparing at the collection level. "Every good URL is an eventual dead link", the authors say, on the basis that a well-used reference will get modified and moved around.
- "Anchor tag" lead to a paper on multimedia presentation and one on coordination. "automatic collection of metadata" leads to a paper on electronic commerce discussing how to organize and present the fields of data attached to merchandise descriptions, e.g. name of company. This might be useful as ways to classify materials since we only now display rather than process HTML Meta fields. The issues here are the selection of schema based on Warwick and Dublin frameworks.

[Well, one-by-one through the A's takes about 30 minutes, including reading 3 papers] Time to take another approach to find the concepts most related, probably under "navigation" or "hyper" but first I'll do a frequency list and see what concepts are most common to this collection of papers. In a way my goal is to "slice" the proceedings, finding the subcollection of relevance.

Another idea, first identify the personal names in the concept list which usually appear as "Last, First" or "Dr. X" or "I.M. Somebody". Our ConceptBrowser. adds this marking to the Iconovex processing for a reasonably accurate list of names. Using names,

- I found one I recognized that lead to a paper on "Features for Personal Web Page Classification" where the problem is to take classified bookmarks and find similar pages using text analysis. This suggests a new feature for twURL to use with rating, e.g. if a "term" is the basis for rating a paper then perhaps all other papers with term share that rating. T
- he name "Furuta" lead to a paper on "Index-based Hyperlinks" where a browser (tool) user selects a word and sees other documents using terms based around that word. This is like our Concept Browser and twURL's string matching. No experiments are reported on how usable this is. Now to frequency ordering:
- "HTML" is the largest subphrase, covering editors, features, etc.
- An article on "accessibility" lead to a paper "30% accessible: a survey of the UK WWW" which I thought might address dead links but dealt with usability of pages based around HTML features.
- "web" lead to an interesting paper on "WebCutter: Tailored Site Mapping", comparing products. BiC certainly contains the notion of browsing a whole site, seeing its overall structure or finding specific content.
- Another mechanisms was presented in "WebTagger: A Bookmarking Service for Organizing URLs" which attempts to address many of the problems of browser-based, individual bookmarking tools with a shared, more descriptive system. twURL could be used for such a system by submitting files of URLs to be shared, retaining identity of submitter.
- "WebSQL: Applications of a Web Query Language" is a paper modeling the web as a relational database over document attributes with extensions to handle path problems (cyclic document references). Such a query language could be applied to twURL's alternative views. By now, I had covered most of the papers and was cycling through them or into papers on irrelevant server technology.
- The final paper under "hyper information content" as "The Quest for Correct Information on the Web" which proposes information formulae related to in- and out-links superimposed upon search engine relevance rankings. Of all the papers reviewed. the last was most relevant to the "information foraging" model and to our own work.

Going back to the Table of Contents, I found a highly relevant paper that hadn't been indexed (a constant problem of our web process is managing file sets where the download or component functions may fail).

- "ParaSite: Mining Structural Information on the Web", Ellen Spertus) bases a number of heuristics on the observation that web text is far more than hypertext in the structuring information carried by URLs while it is also far less than hypertext in that authorship principles are haphazard, e.g. what does a link signify and where is a link drawn. She observes that many inferences can be drawn from considering the nature of links to and from documents, e.g. they're probably on related subjects, and from closure on paths, e.g. prominence of sites followed from different subject areas. With operations based upon many such heuristics, ParaSite (apparently the author's doctoral thesis in progress) should yield results that might be incorporated into the BiC model. The composite of several such systems (WebSQL, Parasite, and twURL) would represent a URL mining system built upon current search engines to server sophisticated web users just as today's information officers use spreahsheets and databases to model and calculate properties of their organizations.

This sample of papers suggests that (1) information driven approaches to web use are still in their infancy with little data to identify better or worse approaches, (2) more work is related to servers, networks, and multimedia than information content. We found several formalizations of part of our work and ideas for several extensions.

Related Conferences

We followed a similar process with several other conferences: Digital Library 95, Hypertext 96, the Allterton 97 Digital Publishing workshop.

Hypertext97 contains a number of papers that address models of documents and document collections. We were unable to access these documents until late in the project but it is clear that new ways of expressing models of navigation are being developed by the hypertext research community. A formal model of "Browsing in Context" could be developed; indeed, the set model underlying LodeStar informally follows that model. There would be several benefits to such a model: (1) better design of the database, user inferface, and editing operations; (2) better explanations in documentation or other tutorials; and (3) a base for more algorithmic analyses.

Other Research Sources

Using the query "information foraging" on Northen Lights Search (this wasn't in other search engines), we found an interesting master's thesis from University of Toronto:

> Abrams, David (1997). Human Factors of Personal Web Information Spaces. Knowledge Media Design Institute Technical Report #1, University of Toronto.

The empirical subject is bookmarks and how people use them, enumerating many problems from a survey. However, the thesis is an encyclopedia of web uses, including many useful categorizations of web practice and empirical data. *Highly recommended.*

Automatic hyperlinking is a new field combining information retrieval and document structuring. For example, "Does Navigation Require More Than One Compass" by Daniela Rus and James Allen shows ways that semantic links may be drawn within a corpora of documents. In a sense, that's what our concept browser does - all documents containing a specific term are linked together. But the author's proposed approach drives more toward ways of permanently linking documents with more semantic content than simple use of terms. One of the surprising keys to making this approach work is ascribing more meaning to document parts, e.g. labelling theorems as "theorem", a practice common to older word processors such as Scribe and LaTex. The emerging technology XML, derived from SGML, is likely to provide many of these opportunities to put structure labelling back into documents, associating structure with appearance separately.

"Category Translation on the Web: Learning to Understand Information on the Internet" (Perkowitz, Etzioni) discusses probabilistic means of relating documents, particularly suitable for very large databases or highly structured data such as email addresses.

IJCAI-95 Workshop on Context in Natural Language Processing lays out a set of questions addressing the nebulous concept of context, e.g. how do you tell whether one context is the same as another. As natural language processing evolves, it may be possible to identify significant contexts, such as industry (automotive vs. medical), depth of document (advertising vs. technical report), or scope of authorship (personal bookmarks V.S. group authored white paper V.S. prize-winning paper).

An extended collection of papers on "Usability on the Web"(International Journal of Human Computer studies) discusses several experiments. An example of serious empirical studies is Tauscher, L. and Greenberg, S. How people revisit web pages: empirical findings and implications for the design of history systems . They point out the great confusion of browser users on what history is, what URLs are being stacked and what forward-backward mean. Since alternative history mechanisms are not likely to replace the fossilized features of modern browsers, it's not clear how the resuls might be used except in add-ons or other browser utilities. Existing packages HindSite and Zooworks package all pages passing trhough a

browser and then index the pages for retrieval by URL or phrase. Other packages accelerate browsers by pre-loading pages permitting more flexible browsing within clusters of pages. We observe, however, that the simplest mechanism of high utility is simply to keep an index of all pages passing into a browser cache as a separate file. A former version of Microsoft Internet Explorer did this, making our LodeStar tools especially useful for processing these histories. Again web utilities exist that index browser's caches. Our concern is that studies like this are focused on features that are already known to be deficient, that what is learned is hard to apply within the current 2-browser marketplace, and that generally we seem to be moving away from the big picture (such as the information foraging model) into interface details and naive human behavior.

The book Secrets of the Super Net Searchers by Reva Basch contains intervuews with numerous Internet figures and experts in the practice of using online databases such as Dialog and Knight Ridder. In general, those librarians and information brokers with many years of experience have higher expectations of search engines because of their past experience with precise, tuned, documented query interfaces (usually text). The allusion to "Gray Literature" (publicity, products, people information) is frequent as a contrast with traditional information types (meeting bibliographic standards) and a distinctive feature of the Internet. Studies of how these experienced professionals use these commercially tuned databases have also shown that only about 20% of the available information in the Internet is actually accessed during queries. While there will always be time limits (and for these professionals also $100/hr online charges), there does not appear to be the packaging tools for collections that we are discussing here. One further concern of these 20 information professionals is that quality is especially a problem on the Internet, constrasting with the commercial databases which were more likely to have certified and regular sources of information. Reading these articles brings an awareness of the near randomness of the web in contrast with current physical library and commercially packaged databases.

To summarize, there is considerable literature on navigational models, on information space characteristics, and on user behavior. We found no model exactly like ours but rather that our approach overlaps many. We believe that the domain-link-concepts combination is distinctive because it draws out clues from different types of sources. In addition, LodeStar presents these views in a limited but familar and consistent interface with myriad operations. Particularly beneficial to add would be: more graphical visualizations, additional text analyses (e.g. document typing and structure extraction), and a clearer browsing model combining our views with traditional one-URL-at-a-timebrowsing.

Web Patterns - Capturing Requirements and Knowledge

We will be using the formulation of the "design pattern" or "pattern language" R&D community, now also being applied to Internet situations (CHI 97 workshop). to identify our assumptions and rationale for the "browsing in context" model On the one hand the pattern approach is quite liberal, permitting various characterizations of essentially problem-solution-consequences. We adopt this approach for several reasons:

1. It forces us to "axiomatize" our theory and reasoning behind the BiC model.
2. The formulation of web patterns, i.e. problems and solutions in a WWW setting, may increase understanding by identifying common problems or solutions to problems. These may range from groupware concerns to detailed algorithms or system architectures.
3. Professionally, we may find it possible to attract more interest in and understanding of our work by riding on a movement with both industrial and academic ties.

Pattern #1 : Defining Web Contexts

Setting: Web users access a wide variety of materials from multiple sites across millions of URLs, using search engines, crafted guides, and personal/professional knowledge. For many technical topics, 1000s of URLs are candidates of interest for typical web tasks.

Problem 1a: An unordered list of URLs is difficult to browse.

Problem 1b: It is difficult to understand the patterns and trends of a subject from an unordered list of URLs.

Solution: Find a meaningful ordering or classification scheme over the URL space and use it to sort the URLs into classes with the same assignment or into order. Classes and ordering proximity should indicate some commonality or distinctive difference among grouped or ordered URLs.

Consequences 1a: Browsing the list in the given order will provide the browser (person) with some sense that one URL following another are likely to either (a) have something in common or (b) be distinguished from each other in some known way.

Consequences 1b: Orderings provide ways of seeing extremes, distributions, centers, and gaps. Comparing ordering provides ways of distinguish collections and identifying time and other trends.

Implementation 1.1: Order URLs by number of links to each within a given pool or links. Extract the pool of links from a set of web pages.

Example 1.1.a. The pool may be generated from a pages generated by web search engines. Then # links to a URL indicates, literally, whether multiple search engines have indexed the page under exactly the same URL.

Forces 1.1.a.1 Webmasters, web page authors, and independent vendors often prime the search engines by explicitly submitting their URLs and by defining HTML Meta information or prefaces with frequent or chosen terms in order to place their web pages highest in a given search engine's relevance scheme. Promotional efforts may influence a URL's position in the ordering.

Forces 1.1.a.2 Search engines may index or revisit only part of a site and miss pages of relevance. Global web processes may influence a URL's frequency of linking in the ordering.

Forces 1.1.a.3 URLs with longer longevity are likelier to have references within the search engines or crafted lists.

Implication 1.1. A context is established for each URL within the ordering, depending upon interpretation as popularity, promotional effectiveness, prominence of organization, longevity of web presence, etc. Specific interpretations may be: a "1-link" URL may indicate a very specific, potentially useful, piece of information; a high-link URL may be a clearinghouse or a highly influential document.

Implementation 1.2: Order URLs by Internet names.

Example 1.2.a (e.g. .edu, .uk), sites (e.g. stanford.edu, ca.ac.uk), server (...) ordering alphabetically within each level.

Forces: Internic domain name assignment follows a rationale of type of organization requesting a name, where the name usually symbolizes some aspect, or its name itself. Additional domain names may in the future be more specific about type of business or purpose of web pages.

Implication 1.2. Browsing or evaluating URLs by domain context can utilize the domain naming rationale, e.g. .edu as an academic institution or association, .com as a company or major network, .net as an Internet provider, ...

Implementation 1.3: Extract phrases that, in a natural language sense of context, correspond to objects, persons, places, events, or concepts.

Order all phrases alphabetically and separate apparent person names from other phrases. For each phrase show each page containing it and a few surrounding words.

Forces: Concept analysis is a difficult text analysis process, requiring some grammatical and dictionary mechanisms. Approximate techniques may assist or may be confusing depending on their precision and completeness.

Implication 1.3: Phrase listings may help the browser/evaluator to understand the scope and nature of the domain. Phrase usage may indicate whether a page is within the topic or not, e.g. "no-fault" is not a part of the "fault tolerance" subject. Phrase profiles may help characterize a document during a scanning phase. Identical phrase lists may indicate identical pages. Modifications in phrase lists on different versions of a page may indicate significant changes.

Pattern Family #2: Bookmarks

Definition: Bookmark file is a collection of URLs selected for frequent access directly from the user's browser.

Context: WWW users frequently visit the same URLs over and over and need an online list for rapid lookup and direct link from the list to the WWW.

Forces/Constraints: Heavy WWW users need to manage large collections.

Problems: (from Tauscher et al)

1. Avoiding long lists by establishing bookmark hierarchies
2. Deciding where to put a bookmark in an existing hierarchy
3. Reorganizing a hierarchy of bookmarks
4. Managing multiple bookmark files
5. Annotating bookmarks
6. Tracking changes in bookmarks
7. Building and managing very large bookmark (>50) collections 8.

Pattern Family #3: Finding Common URLs in Multiple Lists

Setting: Web users build up long lists of bookmarks or collect files with long lists. They often want to determine if a URL is multiply referenced (maybe it's more important than others or interlinks fields)

Problem: Given two (or more) lists of URLs, identify the ones on both lists.

Solution 1: Go down the first list and check off each URL that is on the second list. Delete the ones not on both lists or recopy the common ones to a separate list

Solution 2: Order the lists and merge them attaching a count where one URL merges into another. Extract from this list the ones with counts greater than one.

Consequences: Neither of these solutions is practical for manual use where the lists are more than a dozen.

Implementation 1: Classify URLs into buckets for 1, 2, ... For a given URL, find its bucket, say n, then move it to n+1, adding it to bucket 1 if not in any other buckets.

Implementation 2: Classify URLs by URL paths, attaching an identity for the list to the URL leaves.

Discussion: When are 2 URLs the same? Textual identity is one measure, but URLs may be written with or without port numbers, omitting default files for directories, include anchors within a page. Furthermore, web servers often redirect from one server to another so there is no canonical address for a URL (so it can be moved). Also, many international mirrors exist: is it significant to count both the page in Australia and the page in the U.S.?

This is only a start at defining some of our experience using the "browsing in context" methodology. We find the patterns approach conducive to clearer statements of the purpose, process, and pitfalls of the approach, but are not sure of the value for others learning about the approach.

Conclusions and Further Work on Browsing in Context

During this project, we have begun to get a grasp on some of the empirical characteristics of the web information space, in the spirit of "information foraging" but without the characterizing formulae. We have used tools, described in Appendix C and demonstrated at ONR on September 17, 1997. We have experimented with collecting and portraying topic collections of web pages using three complementary views: domain, links, and concepts. We have defined a canonical set of tasks in a typical professional web use setting and compiled the data for evaluating user response. And we have used web patterns to begin to define some of the ways that common problems occurring during web use may be addressed by "browsing in context". User-driven tools have helped us perform our empirical investigations and codified our model.

Understanding how to use the WWW productively is a big challenge today for all individual information professionals and all information-intensive organizations. In a sense, for heavy-duty information users, the web browser technology has gone the wrong way. Histories, bookmarks, and other aides have become fossilized within dominant browsers that address multimedia, advertising delivery, and operating systems rather than the original objective for the web more like a large technical library. The nature of the marketplace favoring browser vendors who control the rendering of HTML has made it difficult for tools supporting alternative uses of the web to mature. The scale of the web and the ability of enough people to "just about cope" with that scale, together with the genuine opportunities to access both traditional and "gray" (products, resumes, newsletters, etc.) materials, has both revolutionized modern information and trapped it. More research is needed to define viable alternative and complementary tools and techniques to maintain the quality uses of the web for government, scientific, and educational communities.

The following opportunities could be pursued.

More empirical investigations of the web information space and better dissemination of that information.

It surprises many information professionals to learn the reality of how much (or how little) of the web is indexed, how recently, and what recall/precision mean to a commercial search company. Closer to our goals, very little is known about the size of technical topics, the rate of change of this class of materials, let alone the quality aspects of the material and its authors. A few librarian researchers (an informal session at Online World 97) and the Xerox-Stanford experiments could be complemented by a supported effort to collect, validate, and track web information at the topic level.

We believe that the most benefit can now be gained by tapping into the professionals who are systematizing what reference librarians do for their clients (build good lists of resources), e.g. packaging information on important research programs for program managers, proposal submitters, technology transfer agents, and investors. The best improvements in the long run will come from information science researchers working with domain specialists and end users.

One challenge might be to take a topic of significant interest to an important community and capture ALL there is on the web, organize that to suit defined users, monitor and evolve the collection, and use it as a testbed for research on analysis tools, user interfaces, information retrieval, and cognitive models.

Is this doable? We have assessed the quality and quantity of WWW information on the subject of "software safety" with the target community associated with a forthcoming standard from the Underwriters Laboratory. Based on our FMEA experiments, we estimate some 20,000 URLs covering 100s of research groups, products, standards organizations, and the job market. Would it be worth it? e.g. to the Navy and the safety industry to have such a compilation of materials, further organized to match the professional criteria of engineers, researchers, and the public interested in computer risk? Perhaps, depending upon the costs of the compiling, qualifying, and maintaining the collection. We suggest a combination of domain-specific testbed (to attract users and provide coherence) together with objectives similar to the NIST-directed TREC projects (to compare techniques, drive tool improvement, train analysts, and gauge progress).

This is the kind of alternative vision enabled by an understanding of the value and costs of WWW material together with prototype tools, techniques, and skills to head toward that vision.

More alternatives to one-URL-at-a-time browsing promoted by Mosaic and enforced by Netscape and Internet Explorer.

Clustering and our multiple views provide new on-the-fly information packages that can be used to drive browsers, rather than the other way around. HTML renderers (the heart of the modern browser) should be components of tools that provide users with ways of managing URL collections and, when needed, viewing them.

Our brief foray into the research on hypertext, navigation metaphors, information retrieval, and empirical studies convinces us there are strong alternatives for a next generation of browser technology. While the market forces may prevail for the vast number of current browser users, another type of browser user community could evolve. This one would base its choices on productivity (work accomplished, tradeoffs), quality (qualifying results for specific uses and users, defining criteria for qualification), learning (getting more information, more flexibly), and analysis (identifying trends, finding gaps, getting the big

picture).

With the specific goal of fully utilizing the very large information space of the web, new techniques can come for processing other large information spaces, e.g. major policy issues or international operations. We claim our three views have significant generality: where did it come from? what is it linked to? what concepts does it address? The major lacking view is time (is it present, past, future?)

Behavioral studies of how web users work are just appearing. But the results seem to be directed toward identifying all the problems with ad hoc, fossilized mechanisms such as bookmarks and histories in the current dominant browsers. Defining and performing studies based on new approaches like our "browsing in context" might lead to genuinely new technologies, or training in new ways of using existing tools. Again, the problem is that information-intensive web users are trapped by browsers that don't serve their needs, yet weigh heavily on their desktops. A good set of principles about how users address certain tasks, e.g. "find the authority in field X" could further the empirical understanding of both web information spaces and user behavior.

Specifically, we see several additions to our pragmatically-driven "browsing in context" approach:

1. More formal models of either information space (e.g. subject and web characteristics such as in our example collections) or information processing (e.g. trade-offs on when to filter, features for quicker qualification of web results).
2. A full-blown experiment along the lines of our "security policy in digital libraries" experiment, performed by empirical studies researchers. This could answer questions such as: when is one view much better than anther, how to users mesh views, what happens if a view is dropped or added (e.g. adding time).
3. Technical improvements using other research results, such as comparing documents (identifying mirrors), estimating document type (e.g. personal page vs. technical report), slicing (e.g. expanding the context of a phrase to include related document sections), and subject matter organization (ontology building, argumentation representation, and classification using concepts and domains).

References

Index of URLs in this report organized by Internet Domain and by Report File

Use left/right buttons to link to previous/next Internet domains, file icons to go to top level.

by-Internet-Domain

Lower com { 25 }, edu { 9 }, mil { 1 }, net { 1 }, org { 1 }, other { 13 }

com { 25 }

Lower apple { 1 }, ciolek { 1 }, digital { 1 }, discovery { 1 }, ferretsoft { 1 }, ffg { 1 }, hotbot { 1 }, iconovex { 1 }, infoseek { 1 }, lincom-asg { 1 }, metacrawler { 1 }, nytimes { 1 }, onlineinc { 1 }, roir { 7 }, surflogic { 1 }, tympani { 1 }, useit { 1 }, winzip { 1 }, zooworks { 1 }

- apple { 1 } www.atg / personal / tom_erickson /
  - **PatternsWorkshop.html; applied to Internet situations (CHI 97 workshop)**
    http://www.atg.apple.com/personal/Tom_Erickson/PatternsWorkshop.html
- ciolek { 1 } www /
  - **NFORMATION QUALITY WWW Virtual Library**
    http://www.ciolek.com/WWWVL-InfoQuality.html
- digital { 1 }
  - **Altavista**
    http://altavista.digital.com
- discovery { 1 } izzy.online / dco / doc / 1012 / world / nature / canopy /
  - **Discoveryarticle;**
    http://izzy.online.discovery.com/DCO/doc/1012/world/nature/canopy/canopyopener.html
- ferretsoft { 1 }
  - **WebFerret**
    http://www.ferretsoft.com
- ffg { 1 }
  - **WebWhacker/WebSeeker**
    http://www.ffg.com
- hotbot { 1 }
  - **HotBot**
    http://www.hotbot.com
- iconovex { 1 }
  - **EchoSearch; Iconovex Corporation**
    http://www.iconovex.com
- infoseek { 1 }
  - **Infoseek**
    http://www.infoseek.com
- lincom-asg { 1 }
  - **OwWwl.**
    http://www.owwwl.lincom-asg.com
- metacrawler { 1 }
  - **Metacrawler**

http://www.metacrawler.com

- nytimes { 1 }
  - **New York Times travel article**
    http://www.nytimes.com

- onlineinc { 1 } www / pempress / super /
  - **Secrets of the Super Net Searchers by Reva Basch**
    http://www.onlineinc.com/pempress/super/index.html

- roir { 7 }
  - www **twURL; Research Outlet and Integration website**
    http://www.roir.com
    - **authors of this survey (Susan Gerhart and Ted Ralston)**
      http://www.roir.com/bios.htm
    - **"context-drivenbrowsing"**
      http://www.roir.com/chi-browsing-in-context.htm
    - **http://www.roir.com/context-survey.htm**
      http://www.roir.com/context-survey.htm
    - **analysed**
      http://www.roir.com/therac.htm
    - **twURL**
      http://www.roir.com/whatis.htm
    - year2000
      - **programmingresearch**
        http://www.roir.com/Year2000/year2000.htm

- surflogic { 1 }
  - **Surfbot**
    http://www.surflogic.com

- tympani { 1 }
  - **NetAttache Pro**
    http://www.tympani.com

- useit { 1 } www /
  - **CHI97 Workshopon Augmented Conceptual Analysis of the Web, March 1997, Atlanta GA.;**
    http://www.useit.com/chi97/

- winzip { 1 }
  - **Winzip**
    http://www.winzip.com

- zooworks { 1 }
  - **Zooworks**
    http://www.zooworks.com

edu { 9 }

Lower gmu { 1 }, si { 1 }, tamu { 1 }, toronto { 1 }, uiuc { 2 }, umass { 1 }, unc { 1 }, wayne { 1 }

- gmu { 1 } www.isse / jiis /
  - **"Category Translationon the Web: Learning to Understand Information on the Internet"(Perkowitz, Etzioni)**
    http://www.isse.gmu.edu/JIIS/index.html

- si { 1 } photo2 / crane /
  - **Smithsonian pictures**
    http://photo2.si.edu/crane/crane.html

- tamu { 1 } csdl / dl95 /
  - **DigitalLibrary 95**

http://csdl.tamu.edu/DL95/contents.html

- toronto { 1 } www.dgp /
  - **Abrams, David (1997).Human Factors of Personal Web Information Spaces.**
    http://www.dgp.toronto.edu/~abrams/

- uiuc { 2 }
  - chip.cs users / patterns /
    - **"designpattern" or "pattern language" R&D community**
      http://chip.cs.uiuc.edu/users/patterns/patternsold.html
  - edfu.lis allerton / 96 /
    - **Allterton97 Digital Publishing workshop**
      http://edfu.lis.uiuc.edu/allerton/96/wrap_up.html

- umass { 1 } hobart.cs / ~allan / papers /
  - **"DoesNavigation Require More Than One Compass" by Daniela Rus andJames Allen**
    http://hobart.cs.umass.edu/~allan/Papers/aaai-f95-abstract.html

- unc { 1 } www.cs / ~barman / ht96 /
  - **Hypertext96**
    http://www.cs.unc.edu/~barman/HT96/index.html

- wayne { 1 } www.cs /
  - **IJCAI-95 Workshop on Contextin Natural Language Processing**
    http://www.cs.wayne.edu/context/

mil { 1 }

navy { 1 } / www.onr / sci_tech /

- **ProgramManager**
  http://www.onr.navy.mil/sci_tech/personnel/

net { 1 }

imt { 1 } / cu / ~notess /

- **columns and tutorials**
  http://cu.imt.net/~notess/search/

org { 1 }

dig-lib { 1 } /

- **Digital Library Newsletter**
  http://www.dig-lib.org

other { 13 }

Lower cl { 1 }, g { 9 }, om { 1 }, uk { 2 }

- cl { 1 } puc / aroi.ing / ~dfuller /
  - **IBIS**
    http://aroi.ing.puc.cl/~dfuller/conklin.html

- g { 9 } roi / onr /

- final report
  - **Browsing in Context - A perspective on conceptual analysis of the web**
    g:\roi\onr\final report\chi-browsing-in-context.htm
    Links FROM:
    1. "CHI97 Workshopon Augmented Conceptual Analysis of the Web, March 1997, Atlanta GA." < http://www.useit.com/chi97/
    2. "Surfbot" < http://www.surflogic.om
    3. "NetAttache Pro" < http://www.tympani.com
    4. "WebWhacker/WebSeeker" < http://www.ffg.com
    5. "EchoSearch" < http://www.iconovex.com
    6. "twURL" < http://www.roir.com
    7. "analysed" < http://www.roir.com/therac.htm
    8. "programmingresearch" < http://www.roir.com/Year2000/year2000.htm
  - **g:\roi\onr\finalreport\context-survey.htm**
    g:\roi\onr\final report\context-survey.htm
    Links FROM:
    1. ""context-drivenbrowsing"" < http://www.roir.com/chi-browsing-in-context.htm
    2. "NewYork Times travel article" < http://www.nytimes.com
    3. "Discoveryarticle" < http://izzy.online.discovery.com/DCO/doc/1012/world/nature/canopy/canopyopener.html
    4. "Smithsonianpictures" < http://photo2.si.edu/crane/crane.html
    5. "twURL" < http://www.roir.com/whatis.htm
    6. "modelof browsing and organization" < http://www.roir.com/chi-browsing-in-context.htm
    7. "computer-generated "concept"analysis" < http://www.iconovex.com
    8. "packageof supplied reports (1.5MB download)" < http://www.roir.com/context-survey.zip
    9. "roir@netropolis.net" < mailto: roir@netropolis.net
    10. "1.5MBcollection of HTML files." < http://www.roir.com/context-survey.zip
    11. "roir@netropolis.net" < mailto:roir@netropolis.net
    12. "authors of thissurvey (Susan Gerhart and Ted Ralston)" < http://www.roir.com/bios.htm
    13. "Research Outlet and Integration website." < http://www.roir.com
    14. "ProgramManager" < http://www.onr.navy.mil/sci_tech/personnel/
    15. "Dr. Helen Gigley" < mailto: gigley@itd.nrl.navy.mil
    16. "http://www.roir.com/context-survey.htm" < http://www.roir.com/context-survey.htm
  - **Browsing in Context - Empirical Baseline**
    g:\roi\onr\final report\empirical.htm
    Links FROM:
    1. "UCSDProfessor Phil Agre's Red Rock Eater List. " < mailto: rre-help@weber.ucsd.edu
    2. "I" < mailto: rre-help@weber.ucsd.edu
    3. "NFORMATION QUALITYWWW Virtual Library" < http://www.ciolek.com/WWWVL-InfoQuality.html
  - **Browing in Context - An Information Foraging Model**
    g:\roi\onr\final report\foraging.html
    Links FROM:
    1. "IBIS" < http://aroi.ing.puc.cl/~dfuller/conklin.html
  - **Instructions for using Browsing in Context materials**
    g:\roi\onr\final report\instructions.htm
    Links FROM:
    1. "roir@netropolis.net" < mailto:roir@netropolis.net
    2. "suppliedmaterials for the survey" < http://www.roir.com/context-survey.zip
    3. "Altavista" < http://altavista.digital.com
    4. "Infoseek" < http://www.infoseek.com
    5. "HotBot" < http://www.hotbot.com
    6. "Metacrawler" < http://www.metacrawler.com
    7. "OwWwl." < http://www.owwwl.lincom-asg.com
    8. "1.5MBsurvey data file" < http://www.roir.com/context-survey.zip
    9. "Research Outletand Integration " < http://www.roir.com
    10. "Winzip" < http://www.winzip.com
    11. "Iconovex Corporation" < http://www.iconovex.com
  - **Browsing in Context - Introduction**
    g:\roi\onr\final report\intro.htm
    Links FROM:
    1. "columns and tutorials" < http://cu.imt.net/~notess/search/
    2. "EchoSearch" < http://www.iconovex.com
    3. "WebFerret" < http://www.ferretsoft.com
    4. "DigitalLibrary Newsletter" < http://www.dig-lib.org

- **Browsing in Context - Literature Review**
  g:\roi\onr\final report\literature.htm
  Links FROM:
  1. "DigitalLibrary 95" < http://csdl.tamu.edu/DL95/contents.html
  2. "Hypertext96" < http://www.cs.unc.edu/~barman/HT96/index.html
  3. "Allterton97 Digital Publishing workshop" < http://edfu.lis.uiuc.edu/allerton/96/wrap_up.html
  4. "Abrams, David (1997).Human Factors of Personal Web Information Spaces." < http://www.dgp.toronto.edu/~abrams/
  5. ""DoesNavigation Require More Than One Compass" by Daniela Rus andJames Allen" < http://hobart.cs.umass.edu/~allan/Papers/aaai-f95-abstract.html
  6. ""Category Translationon the Web: Learning to Understand Information on the Internet"(Perkowitz, Etzioni)" < http://www.isse.gmu.edu/JIIS/index.html
  7. "IJCAI-95 Workshop on Contextin Natural Language Processing" < http://www.cs.wayne.edu/context/
  8. ""Usabilityon the Web"(International Journal of Human Computer studies) " < http://www.hbuk.co.uk/ap/ijhcs/webusability
  9. "Tauscher,L. and Greenberg, S. How people revisit web pages: empirical findings andimplications for the design of history systems " < http://ijhcs.open.ac.uk/tauscher/tauscher.html
  10. "Zooworks" < http://www.zooworks.com
  11. "Secretsof the Super Net Searchers by Reva Basch" < http://www.onlineinc.com/pempress/super/index.html
- **g:\roi\onr\finalreport\lodestar.htm**
  g:\roi\onr\final report\lodestar.htm
  Links FROM:
  1. "WebFerret (http://www.ferretsoft.com)" < http://www.ferretsoft.com
  2. "Surfbot, (http://www.surflogic.com)" < http://www.surflogic.com
  3. "NetAttache Pro (http://www.tympani.com)" < http://www.tympani.com
- **Browsing in Context - Web Patterns**
  g:\roi\onr\final report\patterns.htm
  Links FROM:
  1. ""designpattern" or "pattern language" R&D community" < http://chip.cs.uiuc.edu/users/patterns/patternsold.html
  2. "appliedto Internet situations (CHI 97 workshop)" < http://www.atg.apple.com/personal/Tom_Erickson/PatternsWorkshop.html
  3. "Tauscheret al)" < http://ijhcs.open.ac.uk/tauscher/tauscher.html

- om { 1 } surflogic /
  - **Surfbot**
    http://www.surflogic.com

- uk { 2 }
  - ac ijhcs.open / tauscher /
    - **Tauscher,L. and Greenberg, S. How people revisit web pages:**
      http://ijhcs.open.ac.uk/tauscher/tauscher.html
  - co www.hbuk / ap / ijhcs /
    - **"Usabilityon the Web"(International Journal of Human Computer studies)**
      http://www.hbuk.co.uk/ap/ijhcs/webusability

Appendix A:
CHI 97 Workshop Paper

Browsing In Context

Susan L. Gerhart, ROI Joint Venture slger@netropolis.net

Position Paper for CHI97 Workshop on Augmented Conceptual Analysis of the Web, March 1997, Atlanta GA.

Perspective

Let's say I'm a "web information professional" who regularly needs to collect extensive amounts of technical material on particular web topics for specific client needs, catalogs "on spec", or personal growth. Search engines and pathfinder pages will typically bring in several hundred URLs -- an eclectic mess of diverse content, relative and absolute quality, and accessibility. What techniques do I employ? What tools do I assemble? What conceptual framework do I use to trade-off quantity and quality for specific projects and for evolution of my intellectual property base? Also, I need to turn my search results and polished reports around quickly, both to meet deadlines and to be cost-effective.

Phenomena Observed

Following are several gross empirical claims based on experience with topics such as those in the attached table of data:

1. Typically, 10-20% of URLs from just about any source (search engines, processed lists, websites) are erroneously constructed, dead, transiently unavailable, etc.
   ==> Large scale topic collection requires specific screening for dead links.
   ==> Internet congestion and unpredictability requires replay and incremental strategies.
2. Running the same query through multiple search engines, even down to depths of 100, will rarely produce very many overlapping URLs.
   ==> Multiple searchers and meta searchers must be plied and then their results collectively collated. Hits from multiple search engines increases likelihood of relevance (because of different retrieval methods) but appearance in only one does not correlate highly with relevance.
3. Many topics tend to be concentrated in specific Internet domains (and sites) but most are also spread across almost all domains (even .mil). Likewise most topics have contributions ranging from individuals through organized entities (with supported websites) through multi-institutional associations. Domain and organization units are only two or several dimensions of web contexts.
   ==> browsing by context can reduce confusion and reveal, sometimes surprising, patterns of content.
4. Rule of thumb: Half the URLs discovered will be keepers, half will be redundant, dull, off-topic, or of very little content.
   ==> Rapid decision making and filtering must be supported, with minimal overhead in locating, downloading, and managing storage of web materials.

Discussion

Several months of web foraging and increasingly disciplined topic organization have revealed the pros and cons of different ways of working the web. On-line surfing is needed and useful only to access initial topic pointers, (typically search engines, clearinghouses, and personal knowledge) and then later follow up leads. Not only is one-URL-at-a-time click-and-wait mind-numbing but also offers the temptation and distraction from the goal of the topic collection. Downloaders such as (Windows 95 products) Surfbot, NetAttache Pro, WebWhacker/WebSeeker, EchoSearch, and our twURL, etc. can collect the initial content for offline browsing (depending of course, on open representation of downloaded "data bases").

But that's only the start. Recording impressions, classifying URLs, and generating reports are the new tasks that absorb time. List editors, bookmark managers, and other browser accessories are required to manage these tasks. Just as "office suites" of 5 years ago formed from word processing, spell checking, drawing tools, and macro languages, new suites are forming to master these time-consuming tasks. Our experience building twURL is that (1) providing a wide array of recording mechanisms can significantly increase the user's ability to concentrate on reading and assessing content and (2) an embedded, programmable browser, even one moderately-impaired wrt modern multimedia and HTML, provides a significant jump in capability. A recent PC magazine "Abort-Retry-Fail" cartoon shows Ross Perot on Larry King Live saying "enough about the election, let's talk browsers. The two big ones aren't reflecting the needs of users. It's time for a 'reform browser'." Well, maybe the cartoon Perot was right about one thing - the browser warriors might be missing out on the essence of the problem from a content-analyzer's perspective.

Assuming the mechanical and logistic problems were (being) solved, how would this topic collector and organizer and analyst get down to the heart of the problem? Our experience is that a good set of user-defined multi-dimensional classification schemes tailored to the specific analysis goal and topic can provide the focus and decision-making. For example, most URLs, i.e. their pages, will fit into the following "genre" classifications:

Organizational unit:
    individual, project, division, institution, association
Utility:
    Method, tool, process, evaluation, theory
Stakeholder:
    User, vendor, consultant, educator/student, researcher, ...
Perspective:
    Fact, issue/controversy, exploration, exposition, ...
Strength (relative to goals):
    great, good, OK, weak
Disposal:
    dead, dull, dumb, different
Process:
    requirements, design, implementation, validation, documentation, education, maintenance, ...

Now, assuming that (1) downloading, browsing, recording, and reporting are small constant parts of the processing for each URL and that (2) classification schemes such as the above provide frameworks for quality discrimination, as well as useful tailoring for end users of the topic collections, what is the most significant part of the decision-making operation and determiner of how long it takes? Consider the task of making a one-minute assessment of a page sitting in a browser. The decision may be a first cut requiring little reading (e.g. a directory listing or an "under construction") or it may require some reading, scrolling, and thinking. If there is no sequence to the URL progression (perhaps 500 or so to be processed), then the overhead of changing contexts is enormous for most of the diverse web collections - one minute a grad student bookmark list, the next a major research luminary's technical report, then a tools vendor, a conference, etc. The more ways the material can be clustered to provide continuity to the decision-maker's thinking, i.e. knowing that this URL fits into some previous established region of the web, if only .com Vs .edu Vs .mil or assurance of multiple search engine hits, then a major amount of context-switching is reduced. Not only can the analyst remain sane while working at stretches of an hour or so, but the resulting classification may be more consistent and reproducible.

Thus, our take on one of the major needs of web users is for better "browsing in context", not just to perform the above challenging analysis tasks but also for regular browsing. Somewhere along the line, browsers (the big ones needing 'reform') turned into motion picture theaters rather than hypertext navigators. Nothing wrong with the former, but many users (a few million or so) most likely need better tools for exploiting the WWW as a vast technical library and personal growth environment.

Trends and Needs

Several technology and infrastructure improvements are needed:

1. Some way of marking and disposing of once-useful, maybe still useful, but no longer timely content, e.g. past newsletters, conference announcements, course offerings, etc. Not just relocatable URLs, but more graceful conventions for archiving and tracking past events, e.g. differences among course offerings, or follow-up publications from conferences.
2. Automatic categorizations or self-classification of organizational units. Knowing whether you're going to a some sheep doctor's home page or the Human Genome project may matter (or maybe not). The eclectic nature of the web provides both the delight of discovery and learning and frequent unnecessary overhead.
3. Usability engineering of the above-mentioned browser accessories (or complements or energizers) to help design the suites that in the next few years will enable modern researchers tools to match the incomparable expansion of knowledge access through the WWW.

Data

Following is a table of results from several topic collections using our twURL tool - a kind of Swiss Army knife with downloader, outline organizer, editor, report generator, browser, rating recorder, etc. The data shows some of the scale of (still experimental) topic collections, e.g. a filtered 250 up to unfiltered 1250.

The first column describes the topic, how it was collected, and degrees of processing (ongoing). The second column shows the associated domain classification, where "Other" is all international, numeric, and often erroneous URLs. Numbers of form

{X Y} indicate X URLs in that class with Y links from the source of the URLs (either saved or cached search engine pages, downloaded web pages, etc.). Site sizes are shown as X [Y] where Y is the number of sites with Y URLs. Note that these collections typically start out with loads of "search engine junk", including ads, self-refineries, and help.

The 3rd and 4th columns are derived analyses classifying the URLs by # of links, e.g. hit by more than one search engine or other web pages (sometimes multiple links from the same page) and most populous sites in the URL pool. In the case of the COCOMO topic, where every URL was viewed and rated using the one-minute classification regime, also shown is the distribution of URLs across search engines used.

More examples of these topics and discussions of some of the phenomena are included on our website [www.roir.com]. To highlight just a bit of what we learned on each topic,

- "usability engineering" contains very well organized pages, which is a pleasant change
- Therac, an engineering case study, shows enormous eclecticism, ranging from a strong impact on ethics courses to instances of actual use of the device under study (a faulty cancer treatment) to a play about young love in a cancer ward
- Year 2000 is generating plenty of income for website developers although some of the vendor companies have deceased long before Jan 1, 2000 (or is it 2001). More seriously, there didn't seem to be much overlap of interest between this millennial problem and computer science research.
- "Evita" revealed one serious on-line paper about Argentina period history but mostly a lot of Madonna fan clubs around college websites
- COCOMO showed us a cross-section of a market.
- "FASE-VL" is an example of a maintained list of externally supplied references, newsletters, etc.
- FMEA, a very technical topic, has a surprising amount of online material, including a fair number of resumes as well as job opportunities

twURL data from several searches

Usability Engineering Technical, Search engines, Unfiltered(only ads)	com { 109 295 } edu { 67 181 } gov { 11 40 } org { 40 102 } net { 13 33 } mil { 4 9 } other { 163 415 }	•5 { 2 12 } •4 { 27 134 } •3 { 78 298 } •2 { 290 614 } •1 { 8 9 }	•23 [1] org ^acm •21 [1] other ^uk ^ac •13 [2] com ^useit other ^fi ^hut •10 [1] gov ^nist •9 [6]
Therac 25 Software safety case, Search engines, Filtered, annotated, analysed	com { 25 54 } edu { 81 186 } gov { 4 13 } org { 15 37 } net { 2 3 } mil { 1 1 } other { 56 126 }	1 { 71 71 } 2 { 51 102 } 3 { 29 87 } 4 { 14 56 } 5 { 9 45 } 6 { 2 12 } 7 { 2 14 } 8 { 3 24 } 9 { 1 9 }	•1 [90] •2 [10] •3 [8] •4 [1] •5 [1] •16 [1] •24 [1]
Year 2000 Date "Crisis" Vendor lists (August 1996), Filtered programming research	com { 274 553 } edu { 4 4 } gov { 2 2 } org { 2 2 } net { 3 4 } mil { 2 2 } other { 22 31	5 { 3 17 } 4 { 60 289 } 3 { 13 39 } 2 { 1 2 } 1 { 232 281 }	93 [1] 19 [1] 17 [1] 13 [1] 11 [3] 10 [2] 9 [2] 8 [1] 7 [4] 6 [1] 2 [5] 1 [44]
Evita (the movie) (really Madonna) Entertainment, Search engines, Lightly filtered.	com { 69 84 } edu { 17 21 } gov { 0 0 } org { 0 0 } net { 6 8 } mil { 0 0 } other { 34 39 }	•2 { 19 39 } •3 { 1 3 } •4 { 2 8 }	•1 [41] •2 [6] •3 [6] •4 [3] •5 [2] •6 [1] •7 [2] •12 [1]
COCOMO Software estimation tool, Search engines, Fully filtered and rated (COCOMO product)	com { 38 48 } edu { 129 184 } gov { 23 29 } org { 9 10 } net { 4 4 } mil { 9 12 } other { 96 117 }	•infoseek { 59 107 } •metacrawler { 84 141 } •webseeker { 192 274 } •wwwl { 56 65 } •1 { 222 222 } •2 { 72 144 } •3 { 10 30 } •4 { 2 8 }	•1 [65] •2 [23] •3 [9] •4 [4] •5 [5] •6 [2] •7 [1] •9 [2] •10 [2] •23 [1] •47 [1]
Software researchers' website Virtual Library, newsletter,	com { 16 81 } edu { 163 360 } gov { 23 105 }	•6 { 3 24 } •4 { 5 32 } •3 { 116 593 }	•125 [1] •22 [1] •19 [1] •13 [1] •6 [1]

Virtual Library, newsletter, pubs, Website download, Not filtered	org { 3 14 } net { 0 0 } mil { 2 6 } other { 72 294 }	•2 { 56 112 } •1 { 99 99 }	•5 [2] •4 [4] •3 [2] •2 [7] •1 [48]
FMEA Safety technique, Search engines, Unfiltered, (scatter-gather experiment)	com { 372 618 } edu { 174 266 } gov { 69 100 } org { 61 78 } net { 29 41 } mil { 11 15 } other { 488 718 }	•33 { 1 33 } •24 { 1 24 } •6 { 1 6 } •5 { 4 21 } •4 { 18 76 } •3 { 67 207 } •2 { 231 488 } •1 { 881 981 }	•71 [1] •47 [1] •34 [1] •33 [1] •32 [1] •25 [1] •20 [1] •19 [1] •18 [1] •17 [1] •12 [3] •11 [3] •10 [2] •8 [4] •7 [4] •6 [9] •5 [18] •4 [22] •3 [34] •2 [74] •1 [257]

Appendix B-1:
An Experiment on "Context-Driven Browsing"

Dear Colleague,

We're contacting you about participating in a survey experiment we're performing as part of a study on how information-intensive technical professionals use the Web.

The Problem

While "information overload", "lost in hyperspace", and "trying to find a needle in a haystack" describe some of the symptoms, let's hypothetically characterize the problem as lack of "context". At one extreme, we're accustomed to traditional and (increasingly digital) human-ordered library services of metadata, subject classification, indexing, and bibliographies. Contrast this with the sheer size, openness, and diversity of materials available on the WWW and the computational responses of spidering, indexing, and updating URL bases. The key element in using the web productively, we believe, is to establish, as best we can, various "contexts" for searching, organizing, and qualifying the materials we use in our professional work. "Context" must be encoded in the conceptual relationship between the words used for searching and their use in the source texts, but more generally it includes the goals, everyday frameworks, and logical systems of the topic under investigation (as you'll see in our example).

Goals of this Experiment and Survey

The purpose of this survey/experiment is to define what some possible contexts are and how they might be used. This 'experiment' consists of a definition of a set of tasks in scenario mode and the provision of alternative representations. We ask you to describe for us your personal ways of using the web and your reactions to these alternatives. We're not expecting any statistical or quantitative results but, rather, are looking for useful qualitative characterizations of experience. We will pool the results and our analysis for posting on our website (more info on the logistics of the survey below).

This survey is being sent or announced to dozens of information professionals with a broad range of backgrounds: scientists, journalists, information brokers, educators, and webmasters, ... The Office of Naval Research is funding our overall investigation of "context-driven browsing", a new model for user-driven interacting with the web.

The following section contains a role playing scenario. We ask you to adopt the role of a person tasked by their manager to obtain information about a specific information security approach and related technology as part of an evaluation of possible ways to secure sensitive data. The manager requests you to perform certain tasks for which the information you collect from the Web is necessary. We give you a choice of several approaches to use in collecting the information - three we call "traditional" because they are those we believe to be those practiced today, and four we call "new". These four new approaches incorporate information in four representations that we have devised to explore the "browsing in context" idea. We provide you with files that we have processed using our tools to be viewed (offline) in your preferred browser.

Please read the scenario, choose some tasks to perform , and answer the questions at the end. Thank you for your participation.

A Scenario for Web Usage

Memo: **Access Controls in the Canopy Digital Library**
From: Canopy DL Project manager
To: Jack, Jill
CC: other team members

As you know, the Jungle Canopy digital library will be developed and used by many individuals: Biologists, entomologists, ornithologists, pharmacologists, and other other scientists; our DL administrative and research team; and R&D organizations seeking to exploit the knowledge we gain of one of the last uncharted regions of the Earth. We want to retain strong control over the data, both to protect our rights and to ease the administrative burdens on all of us. Assuring that all developers and users have proper access will help maintain the many gigabytes of gathered data, help assure and maintain its validity and integrity, and protect the rights of the data's owners until published and after. We'd like to minimize anybody inadvertently mixing up datasets, or unqualified people labeling or linking to data, and, if it does happen, we'd like traceability to

correct the problem as soon as possible.

At last week's DL forum, I heard about an approach called "access control lists" that's been around in computing for a while but just starting to be applied to Digital Libraries. Jack and Jill, please rev up your web browsers and see what you can find out about this subject and bring your annotated URL list to the next team meeting:

1. **Find a definition of access control in the material available on the web. Identify one specific technique of access control.**
2. **Name one authority in the field of access control?**
3. **Find and qualify consulting or training individuals, firms, or associations we might go to for assistance.**
4. **How would we define and administer use of access control techniques?**
5. **Why have some organizations chosen the access control approach over other modes of authenticating users and approving data uses?**
6. **Find three products that provide access control and the organizations offering products for access control.**

In summary. we need a web survey of the who-what-why-how-where for this subject of "access control" and "security policies", starting our process of "due diligence" to make the right decisions.

See you Thursday, please post the URL list on the Intranet by Tuesday so we can look at it ahead of time and online at the meeting.

By the way, did you see the New York Times travel article a couple of weeks ago showing the jungle top walkways in Belize - that will be one of the projects we're working with. This Discovery article and the Smithsonian pictures help us realize how little is known about the bugs and birds in the treetop ecosphere just because it's so hard to get up there!

Survey Instructions

Now, survey participants, with the above scenario, please address as many of these questions as you choose using the representations described below.

The following list of materials are to be used in completing the six tasks set out above. Use the "traditional" means as you would normally. The "different contextual representations" (items d-g) have been compiled by us using a tool we have developed, called twURL, which implements our model of browsing and organization.

The general idea of this survey is that for each of the above tasks you choose, please use the twURL resources provided and compare the results you get with the ones you'd expect from using the traditional sources.

Our goal is to define the value of the kind of information provided by the twURL model of browsing and information organization. Conceptually, you're operating in a different mode, browsing in a "context" we have established by our processing of downloaded search engine information, in contrast to the unordered (or only organized by site and "relevance") provided directly by search and meta-search engines. Your feedback as experts in this field and experienced web users will help us refine the formats and processes for the tools being developed. The major research question here is: **does it help you to have URLs organized in clusters such as these, so that your browsing takes place in a reasonably well-defined context?**

Traditional Sources

a. Whatever you can find from search engines

b. Clearinghouse, virtual library, or other selected lists

c. Your personal bookmarks or knowledge

Different contextual representations

d. A list of over 1000 URLs pooled from search engines, organized by Internet domain, and annotated by a preface of the first 500 words and any HTML meta content.

e. A list of these URLs ranked by # of links (i.e. multiple hits from different search engines) and another list of ranking for size of sites from the URL pool

f. A computer-generated "concept" analysis -- list of terms and links to pages, organized under key terms of the scenario.

The Survey Questions

You'll need the package of supplied reports (1.5MB download) and further expanation and navigation instructions (provided in the zipped package) to answer these questions. Please respond by email to roir@netropolis.net. Fill in either the HTML version (preferred, e.g. using NetScape Gold or other HTML editor) or use the text version.

General

1. Please describe your background related to the subject area. How long have you used the Web? Have you conducted previous web searches related to this topic? Topics of equal or greater complexity?
2. Please describe your WWW experience; metrics would be helpful, e.g. # bookmarks, MB of saved files, hours per week online,
3. Please describe your general approach to using the Web. Do you plan ahead or is your searching ad hoc? Do you organize your search terms according to any particular rules?

Task-oriented Questions

Please answer the following questions for each of the above tasks you chose using the following template. Just omit the parts you skipped.

- How would you approach these questions using your own practices? What do you feel are the pros and cons of the way you currently work on web tasks like this?
- Please indicate which of the supplied "browsing in context" materials you tried? (items **d-f**)
- Did you feel that the supplied materials offered you any different sense of control or greater context to search?
- What was it about the materials that helped? What was hard to use or understand?
- What would be the ideal way of approaching this task?

Finally, do you have any thoughts on other aspects of the "browsing in context" hypothesis and how a survey/experiment like this might further investigate it.

Summary of tasks and questions (can be editted in HTML):

1. Definition/ Technizue	• .Current approaches • Which non-traditional materials used? • Value of materials? • Advantages/Disadvantages? • Ideal
2. Authority	• .Current approaches • Which non-traditional materials used? • Value of materials? • Advantages/Disadvantages? • Ideal
3. External Advice	• .Current approaches • Which non-traditional materials used? • Value of materials? • Advantages/Disadvantages? • Ideal
4. Policy Definition & Use	• .Current approaches • Which non-traditional materials used? • Value of materials? • Advantages/Disadvantages? • Ideal
5. Rationale	• .Current approaches • Which non-traditional materials used? • Value of materials? • Advantages/Disadvantages? • Ideal
6. Products/	• .Current approaches • Which non-traditional materials used? • Value of materials?

Services	•	Advantages/Disadvantages?
	•	Ideal

To obtain the survey materials and instructions, download the 1.5MB collection of HTML files.

Return the Survey by email to roir@netropolis.net

For more information on twURL (the tool suite used to produce the survey materials), other twURLed case studies, and the authors of this survey (Susan Gerhart and Ted Ralston), please vist the Research Outlet and Integration website.

Survey funded by Office of Naval Research under Contract N00014-97-C-0108, Program Manager Dr. Helen Gigley

---

Posted July 29, 1997 SLG as http://www.roir.com/context-survey.htm

Appendix B-2: Instructions for Performing the "Browsing in Context" Experiment

Following, in Q&A form, is the information needed to use the Context Reports for the "Browsing in Context" survey on the subject of *"Access Control and Security Policies in Digital Libraries"*. Survey authors will be standing by to receive questions at roir@netropolis.net. Please don't hesitate to ask questions.

Questions and Answers:

NOTE: These instructions assume you've downloaded the supplied materials for the survey.

Where did this content come from? How much data is there? Has it been filtered?

URLs on the topic of this survey/experiment were retrieved from the usual cast of search engines - Altavista, Infoseek, HotBot - plus two metasearchers - Metacrawler, and OwWwl.

Very little filtering was performed, mostly removal of advertising and search engine self-referrals. Thus the content contains some off-topic URLs, e.g. addressing "Albanian national security", keypads for building access control, and occasional issues of pornography. Most of the references are to HTML files, so other types of content - .text, .ps, .pdf, etc. -- were not fully processed.

More significantly, all URLs were tested for aliveness by downloading the page on Saturday, July 26, so when you click a URL it's much more likely to be there. This process removed over 15% of the URLs reported by search engines that were either known dead or didn't promptly report in on that day.

The resulting URL set is over 1000 URLs, around 24 MB on downloaded files. The input to this survey are extracted text and links, outlines, navigation links and menus, and analyses -- over 6MB in size.

How do I get the data?

The data is 3 sets of files for the 3 different proposed contexts addressed in the survey questions.

Download the 1.5MB survey data file from the Research Outlet and Integration website.

Unzip it, e.g. using Winzip. You'll need roughly 7MB of space if you unzip it all, but you can unzip each of the context file sets as indicated below. This instruction file gives you the main links to all context files and can be reached by the "for more information" link in the report files.

What's a "context" look like?

We've dubbed the contexts: *Domain, Links, Concepts*. Here's what they mean.

Domain

> URLs are organized by ***Internet domain (.com, .edu, etc.)*** with a catch-all category of "other" (mostly international or numeric). Each domain cluster is a list of the sites in that domain, further broken down into hosts within each site. For example, digital.com, uwashington.edu, etc. This context provides an organizational perspective on the URL - where it came from, its project and parent associations, etc. Each URL is annotated with a preface (1st few hundred words) and any HTML meta descriptions or keywords within the file.

Links

URLs are organized by ***(1) number of links TO each URL*** within the pool (e.g. different search engines pointing to the same URL) and by ***(2) size of sites*** (i.e. number of URLs at the same site according to the domain outline above). This perspective provides an approximation to "popularity", as in how well indexed the pages are within the search engines. **WARNING: search engine results vary widely**. Search engines (a) index different parts of sites and (b) use different URL referencing mechanisms as well as (c) different relevance retrieval approaches, hence the numbers are no more than indicators of possible common hits. Each URL is annotated with a list of links TO the URL, including the text used in the document and the source of the link.

Concepts

This perspective extracts ***keywords, or "concepts"*** from the web pages. We've used a technology called SWAPI (Syntactica Web API) from the Iconovex Corporation to extract significant phrases based on grammatical usage, a lexicon of known words, and rules for geographical and personal names. Each URL is annotated with the concept and personal name lists extracted by SWAPI. As you'll see, about 95% accuracy is achieved in meaningfulness of terms; please disregard mistaken concepts or names. URLs have been further clustered by important terms in the "access control", "security policy", "digital library" area.

How do I navigate these reports?

Each report set comes with a menu file (*-m.htm) that has the top 2 levels of the context outlines, linking into the files containing the actual URL information. For example, in the Domains Context, to get to "stanford.edu", click the top "edu" link, then the "stanford" link in the universities list, and you'll be taken to the URLs associated with Stanford.

The top 2 levels of the context files also contain forward-backward-up buttons (left arrow, right arrow, folder icons) that take you to the next/previous node at each level or up to the next higher level. For example, "stanford.edu" links take you to the preceding and following sites in the universities list or up to the .edu level. From the .edu level, you can go to the preceding .com or following .mil domains. Also included are links that go lower into the outlines, e.g. from the "stanford.edu" level to the hosts at Stanford.edu.

Of course, your browser forward-backward buttons also are operational, so you can navigate both within the context outline and on your browser history.

It may take some practice to get the hang of navigating through this dense material using different browsing strategies. It is not our intent within this survey to assess the effectiveness of these navigational mechanisms but rather to focus on better ways of getting at content using context. We do appreciate comments on improving the report formats and navigational mechanisms, however.

There isn't any exotic HTML in these reports, so most browsers should handle the files, but file size (up to 800KB in some cases) may cause slow loading. You may want to adjust font size and window size for easier viewing.

So, exactly what information is available about a URL?

To reiterate, with examples:

Domain

- Title,
- URL,
- (optional) HTML metadata,
- first ~500 words of the page.

URLs are clustered by directory, host, site, and domain.

Links

- Title,
- URL,
- links to each page with text (the underlined part you'd see),
- source (title, here search results are local files).

URLs are clustered by number of links to them. Site names are clustered by number of URLs at that site.

Concepts

- Title,
- URL,

- (optional) list of person's names
- list of concepts.

URLs are clustered (overlapping) by terms contained in the concept list.

How long should this survey take?

About as long as you could spend, but a minimum of an hour to get into the scenario tasks, peruse one of the reports, and answer some of the questions. There's easily enough content here for several hours of browsing; indeed, we hope some of you will find the material itself worth spending some time on.

How do we report back our results?

Fill out the questionnaire at the end of the survey participation form.

We'd prefer receiving HTML formatted results so we can process the data via linking but you'll also find a text form of the survey. Please email back to roir@netropolis.net.

Index of Files

- Survey call and questions (text version)
- Domain Context Menu
- Links Context Menu
- Concepts Context Menu

Posted July 29, 1997 SLG

Appendix C:

LodeStar: Tools for Exploiting Simple Patterns of URL Distribution and Frequency

For more information, see the ROI Website (http://www.roir.com).

We have implemented and integrated a set of rudimentary algorithms that profile collections of Web materials to build a preliminary picture of what's available, possibly answering the search goal immediately without browsing. Technically, the links (<A HREF="URL">link text <A>) are extracted from a collection of HTML files and then the URLs are sorted and merged and classified by WWW domain (.com, .edu, etc.). The premise here is that the texts of the links, read within a WWW domain setting, provide sufficient clues for a person familiar with the subject to make a preliminary judgment of the value of the referenced page. These clues, together with the "hyperlink skeleton" as outlined (in a GUI treeview), provide a kind of canonical representation that (1) can be subjected to other analyses and presentations and (2) within a GUI environment, can be marked up, edited, and traversed by other parts of the tool suite.

The HTML sources may be acquired from previous searches (explicitly saved or retrieved from the browser's cache) or one of the automatic downloaders discussed above. The canonical outline answers such needs as: show all the URLs from commercial, military, or non-US sources but not university or professional associations. Each URL can be looked at in relation to other URLs from the same host and domain as well as the specific sources in the HTML collection.

The outline structures in our GUI tool are editable, so that the above operations yield outlines (trees) that can be trimmed, marked up, counted, and rearranged into various classifications, e.g. "# links>=2" AND "in .com domain" AND "not a search engine company". These edited profiles as plans can now be used for automated, browsing sessions. Furthermore, as URLs are visited, the GUI supports the attachment of (what we call 'webits') notes, ratings or relevance classifications, and marks for further editing.

A more complete description of this GUI, twURL (tm), is available at http://www.roir.com. Briefly, it is Windows 95, written in Visual Basic, using a browser component from Catalyst Development and the Protoview Data Explorer interface component.. twURL operates within a "virtual suite" of tools that assist in downloading HTML materials, monitoring changes, and multi-searching. Specifically we use EchoSearch (http://www.iconovex.com), a multi-searcher that indexes concepts and names of downloaded materials; WebFerret (http://www.ferretsoft.com) collates in-depth, also providing (with 20MB of ZIP Drive space) for indexed text; Surfbot, (http://www.surflogic.com), a general purpose and open downloader and multi-searcher, site mapper, and URL monitor; NetAttache Pro (http://www.tympani.com), another general purpose, but not open, subscription and brief organizer. We use all search engines as much and as deeply (more than just the 10 early hits), but we find the greatest precision and depth in Ultra (http://ultra.infoseek.com), greatest number in Hotbot (http://www.hotbot.com), breadth and classification in Northern Lights (http://www.nlsearch.com). Online meta-searchers such as MetaCrawler (http://www.metacrawler.com) are also useful, although not as extensive or accountable as our desktop multi-searchers.

An extensive range of editing and viewing functions are available: marking for string content, number of links, data attributes (e.g. errors, whether a local file is available, marking within another outline), classifying URLs with these similar peroprties (string content, title, etc.), trimming marked outlines and nodes, copying/pasting/dragging outline parts. HTML reports are generated corresponding to outlines, with menus and interlinks at higher outline levels. Saved versions of the databases are maintained.

Although there are many "offline browsers" and site-downloaders on the market (see the Consummate winsock Apps list), we found it necessary to build our own simpler downloader. Other tools hold their pages in proprietary or require more setup, while providing more management tools. Tracker simply walks a tree, downloading the files, building an index file, while recording errors.

Recently added to this suite is an interface to text analysis tools from Iconoves Corporation (http://www.iconovex.com). Their SWAPI (Syntactica Web Application Programmer Interface) engine uses grammatical patterns to identify phrases corresponding to persons, places, product names, and concepts. A metasearcher, EchoSearch, queries multiple search engines, downloads files, qualifies results for refined search terms, and generates HTML reports. Unfortunately for topics of any size, the HTML reports reach 5MB or more in size, which most browses cannot load and reload fast enough for comfortable use. Our SwapiBrowser transforms SWAPI output into GUI outlines, using the same embedded browser as twURL, permitting browsing the concept lists, seeing contexts of concepts, and linking to those contexts in the files. Additional functions permit some analysis (separation of person's names, counts of contexts, files related by concepts, and concepts related by files).

twURL's underlying formal model is sets, viewing many of the marking and editing operations as special purpose queries over

URL sets and links.

The LodeStar Process

As LodeStar has been developing, we've found ourseleves increasingly in need of a process, not just tools, especially because we are using several interfacing tools and encounter frequent failure of one or more. Here's the outline of the process we currently use:

1. Acquire URLs using search engines, guide pages, or previous collections
2. Extract and Profile URLs
    1. Domain View
    2. Site Size View
3. Download and excerpt context information (Concepts, HTML meta, Prefaces)
    1. ROI JV's LodeStar (twURL, Tracker)
    2. Other vendors' products: Iconovex SWAPI* and EchoSearch*, Surfbot*, etc.
    3. Use the LodeStar ConceptBrowser on Iconovex* output

- Filter out URLs
    - Off-Topic
    - Redundant
- "Browse in Context" of Domain, Link, Concept Views
    - Automatically navigate (VCR style) 100s of URLs
    - Use contexts to add meaning to URLs
    - Rate URLs using your own scheme
- Generate HTML Reports

Appendix D: "Information Foraging" - Internet Domain View

URL Outline

---

com { 9 }

- apta { 1 }
- bellcore { 1 }
- microsoft { 1 }
- useit { 3 }
- webcrawler { 1 }
- webreview { 1 }
- xerox { 1 }

edu { 24 }

- berkeley { 10 }
- gatech { 2 }
- odu { 1 }
- toronto { 7 }
- tufts { 1 }
- uml { 1 }
- unm { 1 }
- wwu { 1 }

org { 20 }

- acm { 18 }
- cnidr { 1 }
- dlib { 1 }

other { 9 }

- 128 { 1 }
- au { 1 }
- de { 1 }
- ie { 1 }
- it { 3 }
- uk { 2 }

Selected URLs for .org Domain

org { 20 }

Lower acm { 18 }, cnidr { 1 }, dlib { 1 }

acm { 18 }

Lower info, info.sigir, siglink, www, www1

- info top / people /
  - **Lecturer James E. Pitkow**
    http://info.acm.org/top/people/pitkow.html

    Preface:: *James E. Pitkow Graphics, Visualization, & Usability Center Georgia Institute of Technology Atlanta, GA 30332-0280 Phone: (404)355-2559 Fax: (404)853-0673 Email: pitkow@cc.gatech.edu Biographical Information James Pitkow is a Graphics, Visualization, & Usability graduate student at Georgia Tech's College of Computing. He graduated Cum Laude in Computer Science Applications in Psychology from the University of Colorado at Boulder in 1993. It was during an undergraduate research...*

    Persons: Pitkow, James || Pitkow, James E.

    Concepts: Assistanceship || Biographical Information || Boulder || Computer Science Applications || Consumer Interests || Cum Laude || Database project || Email || Fax || General Demographics || Georgia Tech ... Tech's College of Computing || Graphics || GVU's WWW User Surveys || HTML || Lecture Topics || Log file || Measurability || NASA's Earth Observing System Distributed Information System || Phone || Psychology || Space Physics || Statistical research || University of Colorado || Usability Center Georgia Institute of Technology Atlanta ... graduate || Usage Patterns

- info.sigir sigchi / chi95 / proceedings /
  - papers
    - **Information Foraging in Information Access Environments**
      http://info.sigir.acm.org/sigchi/chi95/proceedings/papers/ppp_bdy.htm

      Preface:: *Information Foraging in Information Access Environments Peter Pirolli and Stuart Card Xerox Palo Alto Research Center 3333 Coyote Hill Road Palo Alto, CA 94304 pirolli@parc.xerox.com card@parc.xerox.com © ACM Abstract Information foraging theory is an approach to the analysis of human activities involving information access technologies. The theory derives from optimal foraging theory in biology and anthrop...*

      Persons: Pirolli, Peter

      Concepts: Coyote Hill Road || Information Access Environments || Palo Alto ... Alto Research Center

    - **Table of Contents: Papers**
      http://info.sigir.acm.org/sigchi/chi95/proceedings/papers/toc.html

      Preface:: *Table of Contents: Papers Cognitive Models Display Navigation by an Expert Programmer: A Preliminary Model of Memory Erik M. Altmann, Jill H. Larkin, Bonnie E. John, Carnegie Mellon University Predictive Engineering Models Using the EPIC Architecture for a High-Performance Task David E. Kie...*

      Persons: Ahlstrom, Brett || Allport, David || Altmann, Erik M. || Ashworth, Catherine || Baecker, Ronald || Bederson, Benjamin B. || Beirne, Garry || Bellotti, Victoria || Buxton, William || Carey, T. T. || Carr, Rebecca || Carroll, John M. || Chiu, Patrick || Clonts, Joan || Conway, Matthew J. || Cooperstock, Jeremy R. || Dayton, Tom || Diekmann, Barbara || Edmark, John T. || Fels, Sidney || Fishkin, Ken || Fitts, Law || Fitzmaurice, George W. || Foley, James D. || Furnas, George ... George W. || Gershenfeld, Neil || Harper, Richard || Harrison, Beverly L. ... Susan M. || Hayes, Craig || Hinton, Geoffrey || Hix, Deborah || Isaacs, Ellen A. || Ishii, Hiroshi || Jacob, R. J. K. || John, Bonnie E. || Johnson, Jeff A. || Kabbash, Paul || Karat, John || Kieras, David E. || Kline, Richard L. || Knep, Brian || Kohlert, Douglas ... Douglas C. || Kolojejchick, John || Kurlander, David || Kurtenbach, Gordon || Lamping, John || Landay, James A. || Larkin, Jill H. || Levow, Gina-Anne || Lieberman, Henry || Ling, Daniel T. || Lohse, Gerald L. || Mackay, Wendy E. || Mackinlay, D. || MacLean, Allan || Maes, Patti || Maltz, David || Marshall, Catherine C. || Marx, Matt || Mattis, Joe || McConnell, Daniel S. || McKerlie, D. L. || McLellan, S. G. || Meader, David K. || Melle, William van || Mercuri, Rebecca T. || Meyer, David E. || Mitchell, Alex || Moore, Johanna D. || Moran, Thomas P. || Morgan, Pam || Morris, Trevor || Muller, Michael J. || Myers, A. || Narine, Tracy || Nichols, Sarah || Nigay, Laurence || Olsen, Dan R. || Olson, Gary M. ... Judith S. || Paradiso, Joseph A. || Pausch, Randy || Pirolli, Peter || Posner, Ilona || Ritter, Frank E. || Rodriguez, Thomas K. || Roesler, W. || Rosenstein, Mark || Rosson, Mary Beth || Roth, Steven F. || Savidis, Anthony || Sayre, Rick || Schwarz, Heinrich || Sellen, Abigail || Shipman, Frank M.

|| Smith, David Canfield || Stephanidis, Constantine || Stoakley, Richard || Templeman, James N. || Vicente, Kim J. || Wharton, Cathleen || Whittaker, Steve || Wiedenbeck, Susan || Yankelovich, Nicole || Zila, Patti L. || Zimmerman, Thomas G.

Concepts: Advanced Media || AI Laboratory || Akikazu Takeuchi || Alex Paul Conn || Alias Research || Apple Computer || Arizona || AT&T Bell Laboratories || Bristol || Cognitive Analysis ... Models || Colorado || Communication || Computer Science Laboratory || Corporate Research || Corporation || Delft University of Technology || Design Experiences ... Tools || Doree Duncan Seligmann || End-User Training || Ephraim P. Glinert || Georgia Institute of Technology || IBM T. J. Watson Research Center || Industrial Light || Information Access ... of Others || INFORMATION VISUALIZATION || Innovative Interaction I ... Interaction II || Ishantha Lokuge || J. W. van Aalst || Joëlle Coutaz || Kate Ehrlich || Keio University || Koichiro Tanikoshi || Kumiyo Nakakoji || Laboratoire de Génie Informatique || Lotus Development Corporation || Making Choices || Marita Franzke || Mei C. Chuah || Michigan || Microsoft Research || Multimodal Interfaces || Naval Research Laboratory || Nebraska || New Brunswick || Nottingham || NTT Human Interface Laboratory || Octavio Juarez || Pen Interfaces || Pennsylvania || Pittsburgh || Pixar Tom Williams || Psychology Unit || Ramana Rao || Satu S. Parikh || Scott Hudson || Simon Buckingham Shum || Sougata Mukherjea || Suguru Ishizaki || Table of Contents || Taketo Naito || Texas A&M University || Toronto ... William Buxton || U S WEST Communications || University of Guelph || Upendra Shardanand || Usability Analysis || Uwe Malinowski || Vibhu O. Mittal || Virginia ... Polytechnic Institute & State University || Visual Display Techniques || Wisconsin-Eau Claire || Xerox PARC

- siglink sigchi / chi96 / proceedings /
  - papers
    - card
      - **The WebBook and the Web Forager: An Information Workspace for the World-Wide Web**
        http://siglink.acm.org/sigchi/chi96/proceedings/papers/Card/skc1txt.html

        Preface:: *The WebBook and the Web Forager: An Information Workspace for the World-Wide Web Stuart K. Card, George G. Robertson, and William York Xerox Palo Alto Research Center 3333 Coyote Hill Road Palo Alto, California 94304 E-mail: {card | robertson | york}@parc.xerox.com ABSTRACT The World-Wide Web has achieved global connectivity stimulating the transition of computers from knowledge processors to knowledge sources. But the Web and its client software are...*

        Persons: Computer, personal || Computers, personal || D.A. Henderson || Henderson, D. A. || J.D. Mackinlay || J.M. Designing Workscape || Krebs, J. R. || Pirolli, P. ... Peter || Robertson, George G. || S.K. Card || Shafer, D. || York, William

        Concepts: A New Paradigm || A. van Dam || ACM ... CHI 94 ... Conference ... Press ... Transactions || ACM/IFIPS InterCHI 3 Conference || Affordances || Amsterdam || Animation of pages || Ark Interface || BellCore SuperBook Document Browser || Biology || Book Books ... Metaphor ... simulation || Boston || Brown's Intermedia system || Browsing Web || California ... 94304 || Characteriza tion || CHI 94 ... 95 || Commercial products || Communications link ... of ACM || Computer Application Delivery || Computers || Computing Systems || Conference Companion || Cost of Knowledge Characteristic Function ... structure || Cost-structure of information || Cost-structures of information || Coyote Hill Road || Current Web servers || Demo Book || Display of book || Document Lens || DonÕt || E-mail || Ecology || Electronic Book ... documents || Elsevier Science Publishers || Fig. 7 || Fisheye table of contents || Focus Place ... Position || Font scale ... size || Food rate of gain || Forager || Function of time cost || General Magic || Graphics ... capabilities || History mechanism || Home page ... pages ... pages of real-estate brokers ... position || Home-Page Books || Hot List Books || Hotlist ... browser ... notion ... pages || HTML image ... pages || Human Factors || Human-Computer Interaction || Hypertext 93 ... documents || IEEE Computer || Immediate Memory space ... Storage place ... Storage workspace || Information Access Environments ... enrichment ... Foraging ... gain per unit time ... sources || INFORMATION VISUALIZATION || Information Visualizer system || Information-based work || Information-intensive activities || Information-rich world || Integration of animated 3D book || Interactive animation ... use of information || INTERCHI 93 || Intermediate Storage area || Internet ... speeds || Invention make efficient use of display space || Irre sistible affordances || IsnÕt || ItÕs position || Keyword query || L. Representation || Lycos attempt || Magic Link User's Guide || MAYA Design Group || Mouse button || Netscape 1.1N || NetScape's versions of commands || New York || Operational analogues || Palo Alto ... Alto Research Center 3333 Coyote Hill Road ||

Parc.xerox.com || Personal computer ... computers || Physical book ... books || PIC-1000 || Princeton University Press || Processor power || ProcessorsÑword processors || QuickTime version (1.2M) click || R. C. A. Shiner || Reimplement || Related Work || Relative-URL Books || Riffle || Rooms system || S. Card || Scrollbar || Scrollbars || Search Reports || Seattle || SGI Demo Book || Silicon Graphics ... Graphics Iris computer || Sistible affordances || Sony Corporation || SourcesÑportals || Tertiary Place ... Storage area || Tion || Tokyo || Topic Books || Tunable || URLs leads || Use of 2D physical book simulation || User Interface Software || User-Centered Design || UserÕs own limited time || UST 93 || Web ... analysis methods ... browser ... browsers ... entity ... Forager ... Forager workspace ... information ... page ... pages ... problems ... space || WebBook Applications ... Document Lens view ... page flip uses rigid pages || Weblets || Wide-scale connectivity || Workspace ... manager ... sets || Workspaces || World Wide Web || World-Wide Web || Writing Electronic Book || Xerox TabWorks system

- pirolli
  - **Scatter/Gather Browsing Communicates the Topic Structure of a Very Large Text Collection**
    http://siglink.acm.org/sigchi/chi96/proceedings/papers/Pirolli/pp_txt.htm

    Preface:: *Scatter/Gather Browsing Communicates the Topic Structure of a Very Large Text Collection Peter Pirolli, Patricia Schank, Marti Hearst, Christine Diehl Xerox Palo Alto Research Center 3333 Coyote Hill Road Palo Alto, CA 94304, USA pirolli@parc.xerox.com, schank@unix.sri.com, hearst@parc.xerox.com, cdiehl@violet.berkeley.edu ABSTRACT Scatter/Gather is a cluster-based browsing technique for large text collections. Users are presented with automa...*

    Persons: C.R. Fox || Diehl, Christine || Pedersen, Jan || Pirolli, Peter || R.D. Bliek || Salton, G. || Schank, Patricia

    Concepts: Annuam International ACM/SIGIR Conference || Biology || Boston || Butterworth & Co || CHI-95 || Clustering method || Clusterings || Computing Machinery ... Systems || Conceptual Model of Topic Structure || Concrete course of action || Coyote Hill Road || D. Overview of first text retrieval conference || Distribution of ratings ... of relevant documents ... of Relevant Texts || Diversity of topics || Document clusters ... collection ... retrieval tool ... seeds || Easy No ... queries || Ecology of finding resources || Effective Query Terminology ... topic language || Employee family || Equation 1 || Evaluation issues || Figure 1 ... 2 ... 4 || Finally "medicine || Finding Relevant Articles || Groups x Query Difficulty x Phase factorial || Hard No ... queries || Health risks || Human Factors || Information Processing ... retrieval ... retrieval technique ... retrieval tool ... retrieval tools ... Science || Inter-rater agreement || Interactive method || Interface Effects || Jorunal of American Society || Journal of American Society || Keyword queries || Keyword-search retrieval algorithms || Latin square || Law relation || Medium No ... queries || METHODOLOGICAL ISSUES || Miss" || MSE 108.55 ... 128.98 ... 1737.52 ... 2355.77 ... 297.79 ... 75.71 || National Institute of Standards || Navigation || New York || OVERVIEW of EXPERIMENT || P<.001 || PA. ACM || Pairwise document similarity ... similarities || Palo Alto ... Alto Research Center || Partitional || Phase 1 ... 2 ... 4 || Phases 1 ... 2 || Pittsburgh || Practice Effects || Precisions || Probability of events || Psychological Review || Psychometric evaluation || Query Difficulty ... Effects ... topics || Relation of No. Scatter/Gather windows || Relevance Ratings || Relevant documents || Retrieval techniques || S. Card || Scatter/Gather approach ... browser ... button ... cluster windows ... interaction ... interface ... Speeded ... v. SimSearch || Searching behavior || Second Text Retrieval Conference TREC-2 || Semantic space || SG v. SS ... v. SS group difference || SGR condition ... diagrams ... group ... subjects || SGS 2.78 ... condition ... conditions ... diagrams ... diagrams 2.82 || Show Titles button ... Titles Window || Similarity Search || SimSearch technique || Stanford University || Stuart Card || Table 1 ... 3 || Tipster collection || TIPSTER text collection || Topic similarity data ... structure ... structure diagram ... structure of text collection || TREC conference || Ui || VanRijsbergen || Wihout || Xerox PARC

- www
  - sigchi
    - chi95
      - ap
        - **CHI '95 - AP - Papers: Information Access**
          http://www.acm.org/sigchi/chi95/AP/tue3p1.html

Preface:: *Papers Information Access Tuesday, 2:00 - 3:30 Session Chair: Wendy A. Kellogg, IBM T. J. Watson Research Center Discussant: George W. Furnas, Bellcore Information Foraging in Information Access Environments Peter Pirolli, Stuart Card, Xerox PARC TileBars: Visualization of Term Distribution Information in Full-text Information Access Marti A. Hearst, Xerox PARC...*

Persons: Furnas, George W. || Hearst, P M. || Kellogg, Wendy A. || Mackinlay, D. || Pirolli, Peter

Concepts: Distribution information || Full-text Information Access || IBM T. J. Watson Research Center || Information Access ... Access Environments || Ramana Rao || Searching Citation Links

- electronic documnts /
  - **CHI '95 Proceedings - Keyword Index**
    http://www.acm.org/sigchi/chi95/Electronic/documnts/keyword.html

    Preface:: *Keyword Index This index only includes documents from the Paper and Design Briefing sessions. 3D displays Providing Assurances in a Multimedia Interactive Environment Dynamic Stero Displays 3D interaction Virtual Reality on a WIM: Interactive Worlds in Miniature Dynamic Stero Displays Planning-Based Control of Interface Animation 3D interfaces The "Prince" Technique: Fitts' Law and Selection Using Area Cursors Planning-Based Con...*

    Persons: Fitts, Law || Publishing, Electronic

    Concepts: Architecture || Cartography || Citation graphs || Communications || Computer animation || Cooperative Work || Design Briefing sessions || Domain orientation || Education || Egocentric projection || Electronic mail ... markets ... publishing || Ethics || Ethnography || Formative evaluation || Freeform interaction || Garnet || Gestural interfaces || GOMS models || Graphics presentations || Groupware || Haptic input || HCI professional issues || Heuristics || Home automation || Hypertext || Information retrieval || Integration || International monetary fund || Internet || Keyword Index ... search || Lotus notes || Magic lens || Multimodal interactive systems ... Interfaces || Multiscale interfaces || Navigation || Prototyping tools || Psychology of programming || Silk || Software Engineering || Space-Scale Diagrams || Tabworks interface || Telephone || Touchscreen || Video ... conferencing ... images || Visual Languages || ZStep 94

- proceedings papers /
  - **Information Foraging in Information Access Environments**
    http://www.acm.org/sigchi/chi95/proceedings/papers/ppp_bdy.htm

    Preface:: *Information Foraging in Information Access Environments Peter Pirolli and Stuart Card Xerox Palo Alto Research Center 3333 Coyote Hill Road Palo Alto, CA 94304 pirolli@parc.xerox.com card@parc.xerox.com © ACM Abstract Information foraging theory is an approach to the analysis of human activities involving information access technologies. The theory derives from optimal foraging theory in biology and anthrop...*

    Persons: Pirolli, Peter

    Concepts: Coyote Hill Road || Information Access Environments || Palo Alto ... Alto Research Center

- chi96 proceedings / papers / pirolli_2 /
  - **Silk from a Sow's Ear: Extracting Usable Structures from the Web**
    http://www.acm.org/sigchi/chi96/proceedings/papers/Pirolli_2/pp2.html

    Preface:: *Silk from a Sow's Ear: Extracting Usable Structures from the Web Peter Pirolli, James Pitkow, Ramana Rao* Xerox Palo Alto Research Center 3333 Coyote Hill Road Palo Alto, CA 94304, USA email: {pirolli, rao}@parc.xerox.com, pitkow@cc.gatech.edu *authors are ordered alphabetically ABSTRACT In its current implementation, the World-Wide Web lacks much of the explicit structure and strong typing found in many closed hypertext systems. While this property pro...*

Persons: Diehl, C. || Halvorsen, P-K. || Hearst, P M. || Hogg, T. || J.D. Mackinlay || J.O. Pedersen || J.R. Krebs || Kehoe, C. || Milson, R. || P.L. Pirolli || Pirolli, Peter || Pitkow, J. ... James || S.K. Card || Schneiderman, B.

Concepts: A.k.a. || ACM || ACT theory || Activation Algorithm ... flows ... input ... mechanism ... mechanisms ... Nets ... network ... Source Network ... spread || Active Web Pages Found Xerox Home Page || America Online proxy || Anthropology || Artificial Intelligences || Biology attempts || Black Boxes Hard || Bookwise Home Page || Boston || Botafoga algorithm || Business || Butterworth & Co || Capacitor model || Cdepth || CHI-95 || CHI-96 || Communications of ACM || Computer networks || Computing Machinery ... Systems ... Systems CHI-95 || Cone Tree || Connectivity structure || Consumption of valuable information || Content Nodes || Coyote Hill Road || Csim || Cultural knowledge units || DATA SOURCES || Diet problem || Directory browsers || Division home pages General info Search form || Dynamics of activation || E.A.and B.Winterhalder || Ecology ... of information-bearing items || Email || Entry Points ... Precisio Inlinks Outlinks of ty || Evolution of Web Ecologies || Examples of Web pages || Financial reports || Firewalled domains || Firewalls || Food energy || Forager || Graph networks ... node ... representations of strength of association of WWW pages ... structure ... structures ... theoretic algorithms ... theory || Graph-theoretic techniques || Graphical illustration of improvements || Head Nodes || Heuristics || Home page || Hotlists Home || HTML document header || Human Factors ... memory || Hypertext ... collections ... link ... link topology ... link topology of Web locality ... links ... research ... spaces ... structures ... systems ... Transfer Protocol || Image Handling || Indentifying aggregates || Index Nodes || Inductive technologies || Information Foraging theory ... Foraging Theory attempts ... gain ... retrieval || INFORMATION VISUALIZATION || Informavores || Inlink || Inlinks || Intelligence systems || Intelligent Text || Inter-document text similarities ... Text Similarity || Inter-memory associative link structures || Interactive animation ... mechanisms ... structures || Interests of Home Page Visitors || Internet ... Yellow Pages || ISDN Systems || Iterate Equation 2 || Journal of Experimental Psychology || Junk categories || Lawrence Erlbaum Associates || Linear Equation 1 || Lion coming || Lke title || Logarithmic transform || Logs of requested items || Machine learning techniques ... name || Memes || Meta-document vectors || Meta-Information || Navigation ... patterns || Needed WWW pages || New York || Node Type Size Number Number Depth Similari || Object-Oriented Architecture || Optimality of diet ... of WWW interactions || Optimize Foraging Decisions || Organic environment || Outlink features || Outlinks || OVERVIEW of APPROACH || Oxford University Press || Page characteristics ... k Content || Palo Alto ... Alto Research Center || PARC's Map Viewer || Parc.xerox.com || Pathname parts || Personal Home 1 k & 3 || Perspective Wall || Phase transitions || Physical site || Pirolli || Point n Children Children Index || Predict Needed Information || Princeton University Press || Profitabilities || Properties ... of activation networks ... of information evolving || Property of pages || Psychological Review || Pure topology-based methods || Ramana Rao || Rao parc.xerox.com || Reference Head 0.70 Org. Home Page 0.30 || Report Series Ken Fishkin's Public Home Page || Retrieval mechanisms ... of needed information ... of WWW pages ... techniques || Ripper || Rules of mind || S. Card || S.Hudson. Visualizing complex hypermedia networks || Silk || Similarity measure || Source Index || Sow's Ear || Spread of activation || Spreading activation algorithm || Table 1 ... 2 ... 4 ... of Contents || TDB calculation ... full-text retrieval engine || Technological environment || Text Group project overviews ... Retrieval ... similarity matrix || Th column ... row of matrix || Third ACM Conference || Titles of Page Digital Tradition Keywords RXRC Cambridge Technical || Toc || Topological || Topology ... matrix ... networks ... of Web locality || TOPOLOGY-BASED APPROACHES || Traversals || Ty || Typical Web Author || U. Categorizing || Uniform Resource Locator || URLs of nodes || Usable Structures || Usage Statistics || VanRijsbergen || Vj || Web aggregate || WEB CATEGORIZATION || Web group ... groups ... infrastructure ... Journal ... localities ... locality ... page ... Page Feature Vectors ... pages ... server's file system ... spaces ... structure ... structure aggregates || WEB VISUALIZATION || Wj || World-Wide Web || WWW browsers ... page ... page vectors ... pages ... pages of different functional categories ... server ... site ... sites || Xerox Corporation 1994 Form 10-K ... Fact Book ... Home page ... PARC's Digital Library Home Page ... Web ... Web locality ... Web server ... Web space ... WWW server || Xerox's external Web site

- chi97 proceedings / workshop /
  - **CHI 97: Augmented Conceptual Analysis of the Web**
    http://www.acm.org/sigchi/chi97/proceedings/workshop/jn.htm

    Preface:: *CHI 97 Electronic Publications: Workshops Augmented Conceptual Analysis of the Web Wendy A. Kellogg IBM T.J. Watson Research Center P.O.Box 704 Yorktown Heights, NY 10598 USA +1-914-784-7826 kellogg@watson.ibm.com Jakob Nielsen Sun*

Microsystems 2550 Garcia Ave. Mountain View, CA 94043, USA Email: jakob@eng.sun.com Web: http://www.useit.com/ ABSTRACT A workshop at the ACM CHI97 conference on computer-human interaction, Atlanta, GA. This workshop is by invitation only and tak...

Persons: Kellogg, Wendy A. || M.B. Rosson || Mack, R. || Nielsen, J. ... Jakob || Pirolli, Peter || Property, intellectual

Concepts: Atlanta || CHI GI'87 ... workshops || Computing Systems || Conceptual Analysis of Web || CSCW || Electronic Publications || Engines || General Co-Chair of CHI'94 || HTML || Human Factors || IBM || Intellectual property || Internet software || Inventory of significant Web phenomena || L. Marks || Outstanding Technical Achievement awards || Papers co-chair || Proceedings of CHI'95 || Research Staff Member || Rmation-theoretic analysis of cost structure || SIGCHI positions || Software planet || Web ... design ... pages ... phenomena ... technology ... usability || Web's reach || Workshop ... Organizers ... Report || World Wide Web

- sigs
  - sigchi
    - bulletin
      - 1996.2
        - **SIGCHI Bulletin Vol.28 No.2, April 1996: Cognitive Architectures and HCI**
          http://www.acm.org:82/sigs/sigchi/bulletin/1996.2/kirsch.html

          Preface:: *Issue Article Vol.28 No.2, April 1996 Article Issue Cognitive Architectures and HCI Susan S. Kirschenbaum, Wayne D. Gray, Richard M. Young Table of Contents Introduction Represented Architectures HCI Tasks References About the Authors The Cognitive Architectures and Human-Computer Interaction Workshop examined computational cognitive modeling approaches to human-computer interaction issues (HCI). The five major architectures and variations represented were briefly summarized. Part...*

          Persons: Adelson, B. || Authors, Addresses || Bauersfeld, P. || Bennett, J. || C.B. 344 || Dumais, S. || Erlbaum, Lawrence || Gray, Wayne D. || Kirschenbaum, Susan S. || Moran, Thomas P. || Olson, J. || P.G. (1994) ... (1995) || Pirolli, Peter || Young, Richard M

          Concepts: ACM CHI'92 Conference ... CHI'94 Conference ... Press || ACT-R & Soar ... model || ACT-R's conflict resolution mechanism ... rational analysis mechanism ... theoretical mechanisms || Analogical Reasoning || Ann Arbor || Appendix A || Applied Cognition program || Architecture ... bears || Architecture-defined boundaries || Architectures || Artificial Intelligence || Berlin || Boston || Boulder || Cambridge ... CB2 2EF || Chaucer Road || CHI conferences || Cognitive Architectures ... mechanisms ... Models ... phenomena ... Science ... task analysis ... theory || Colorado || Combinatorics problem || Comparing Architecture || Computational Model of Highly Interactive Task ... Model of Skilled Use of Graphical User Interface || Computer application ... Science || Computing Systems || Connectionist architectures || Construction-Integration theory || Cost-of-Knowledge Characteristic Function of information || Cricket Graph || D. W. & Krebs || Data search || Declarative memory chunks || Declarative/procedural memory distinction || Department of Electrical Engineering || Display-based human-computer interaction || Display-Based Systems || Division Newport || E. A. & Winterhalder || Engineering Psychologist || Environmental Contributions || EPIC Tech || Evolutionary Ecology || Experimental Psychology || Fairfax || Franzke (1994) || G. Lynch || George Mason University's Human Factors || Grain size || Graph-drawing program || HCI ... design problems ... issues ... phenomena ... problems ... task ... tasks || Human Behavior ... Factors || Human-computer interaction issues || Human-Computer Interaction journal ... Interaction Workshop || ICS Tech || Instantiations ... of tools || Institute of Cognitive Science || Interface Consistency || International Journal of Human-Computer Studies || Language text comprehension ... text processing system || LTM of queries || M. &

Polson || Mason University || Memory element ... elements || Model Human Processor || MRC APU || NASA Ames Cognitive Modeling Workshop || Naval Undersea Warfare Center ... Undersea Warfare Center Division || New York || Newport || Perceptual motor performance problem || Ph.D. Dissertation || Princeton University Press || Proceedings of CHI 94 Conference ... of Sixteenth Annual Conference of Cognitive Science Society ... of Vienna Conference || Production firings ... per cycle ... system || Psychological Review || Psychology of Human-Computer Interaction || Raven || Research Scientist || Rieman (1993) || Skilled Interaction || Submarine || Table of Contents || Technical Report 94-02 || Teller machine || Two Cognitive Architectures || UK Medical Research Council's Applied Psychology Unit || University of Colorado ... of Michigan || Writing Desk

- 1997.1
  - **SIGCHI Bulletin Vol.29 No.1, January 1997: AVI '96**
    http://www.acm.org:82/sigs/sigchi/bulletin/1997.1/avi96.html

    Preface:: *Issue Article Vol.29 No.1, January 1997 Article Issue AVI '96 - An International Workshop Peter Pirolli The Third International Workshop on Advanced Visual Interfaces, AVI '96, was held May 27–29 in a converted monastery, in the town of Gubbio, in Umbria, Italy. Monks chose these golden hills of central Italy to build their cloisters, so that they could escape the plains of darker times. These hills exude savory foods and hearty wines that deeply satisfy the body so the mind ca...*

    Persons: Alberti, Leon Battista || Costabile, Dr. Maria Francesca || Pirolli, Peter || Santucci, Dr. Giuseppe

    Concepts: ACM Press ... SIGMultimedia || Advanced Visual Interfaces ... Visual Interfaces workshops || Animation techniques || Apennine mountains || Architecture || Associate Professor || AVI 96 ... 96 hosts ... 98 || Bari || Ben Scheiderman's call || Berkeley || Borgo San Sepolcro || Canada || Carnegie Mellon University || Chile || City states || Commune of Gubbio || Communication || Concrete applications || Coverage of works || Coyote Hill Road || Database research || Ecological perspective || Empricial methods || Esprit IDOMENEUS || Esprit's FADIVA || Europe || Festa dates || Generic design problem || Germany || Gran Sasso || Gubbio || Historic procession || Hotel Ai Cappuccini || Human-Computer Interaction || Information Foraging theory || International aspirations ... Workshop || Isabel Cruz of Tufts University || Italian AVI 96 hosts || Italy || Knolwedge content || Learning-to-learn strategies || Mathematics || Medieval dishes ... music ... times || Multimedia interfaces || Navigation || Olive trees || Palio della Balestra cross-bow competition || Palo Alto || PIE formalism || Portugal || Psychology || Renaissance || Roma La Sapienza || Rustic part of Italy || Saint Clare ... Francis || Sweden || Third International Workshop || Tilling of body || Torchlight || Umbrian countryside ... food ... towns || United Kingdom ... States || Universita di Roma || University of California ... of Toronto || Visual architectures ... databases ... interfaces ... Languages ... layout algorithms ... programming techniques || Web Forager || Workshop arrivals ... proceedings || Workspace || World Wide Web || Xerox Palo Alto Research Center ... PARC

- chi95
  - ap
    - **CHI '95 - AP - Papers: Information Access**
      http://www.acm.org:82/sigs/sigchi/chi95/AP/tue3p1.html

      Preface:: *Papers Information Access Tuesday, 2:00 - 3:30 Session Chair: Wendy A. Kellogg, IBM T. J. Watson Research Center Discussant: George W. Furnas, Bellcore Information Foraging in Information Access Environments Peter Pirolli, Stuart Card, Xerox PARC TileBars: Visualization of Term Distribution Information in Full-text Information Access Marti A. Hearst, Xerox PARC...*

      Persons: Furnas, George W. || Hearst, P M. || Kellogg, Wendy A. ||

Mackinlay, D. || Pirolli, Peter

Concepts: Distribution information || Full-text Information Access || IBM T. J. Watson Research Center || Information Access ... Access Environments || Ramana Rao || Searching Citation Links

- electronic
  - documnts
    - **CHI '95 Proceedings - Author Index**
      http://www.acm.org:82/sigs/sigchi/chi95/Electronic/documnts/author.ht

      Preface:: *Author Index This index only includes documents from the Paper and Design Briefing sessions. Ahlstrom, Brett Building Geometry-based Widgets by Example Allport, David Applying Electric Field Sensing to Human-Computer Interfaces Altmann, Erik M. Display Navigation by an Expert Programmer: A Preliminary Model of Memory Anderson, Steve Designing a "Front Panel" for Unix: The Evolution of a Metaphor...*

      Persons: Beth, Mary || Canfield, David || Fitts, Law || Fulton, Jane || Furnas, George W. || Hearst, P M. || Mackinlay, D. || Moran, Thomas P. || Morgan, G. || Olson, J. || Paul, Alex || Pirolli, Peter

      Concepts: Author Index || Design Briefing sessions || Doree Duncan || Ephraim P || Hudson || Juarez || Lennart E || Linn D || Mei C || Roxanne F || Satu S || Shipman III || Simon Buckingham || Vibhu O

    - **CHI '95 Proceedings - Keyword Index**
      http://www.acm.org:82/sigs/sigchi/chi95/Electronic/documnts/keyword.

      Preface:: *Keyword Index This index only includes documents from the Paper and Design Briefing sessions. 3D displays Providing Assurances in a Multimedia Interactive Environment Dynamic Stero Displays 3D interaction Virtual Reality on a WIM: Interactive Worlds in Miniature Dynamic Stero Displays Planning-Based Control of Interface Animation 3D interfaces The "Prince" Technique: Fitts' Law and Selection Using Area Cursors Planning-Based Con...*

      Persons: Fitts, Law || Publishing, Electronic

      Concepts: Architecture || Cartography || Citation graphs || Communications || Computer animation || Cooperative Work || Design Briefing sessions || Domain orientation || Education || Egocentric projection || Electronic mail ... markets ... publishing || Ethics || Ethnography || Formative evaluation || Freeform interaction || Garnet || Gestural interfaces || GOMS models || Graphics presentations || Groupware || Haptic input || HCI professional issues || Heuristics || Home automation || Hypertext || Information retrieval || Integration || International monetary fund || Internet || Keyword Index ... search || Lotus notes || Magic lens || Multimodal interactive systems ... Interfaces || Multiscale interfaces || Navigation || Prototyping tools || Psychology of programming || Silk || Software Engineering || Space-Scale Diagrams || Tabworks interface || Telephone || Touchscreen || Video ... conferencing ... images || Visual Languages || ZStep 94

    - papers
      - **Information Foraging in Information Access Environments**
        http://www.acm.org:82/sigs/sigchi/chi95/Electronic/documnts/p

        Preface:: *Information Foraging in Information Access*

Environments Peter Pirolli and Stuart Card Xerox Palo Alto Research Center 3333 Coyote Hill Road Palo Alto, CA 94304 pirolli@parc.xerox.com card@parc.xerox.com © ACM Abstract Information foraging theory is an approach to the analysis of human activities involving information access technologies. The theory derives from optimal foraging theory in biology and anthrop...

Persons: Pirolli, Peter

Concepts: Coyote Hill Road || Information Access Environments || Palo Alto ... Alto Research Center

- www1 sigs / sigchi /
  - bulletin
    - 1996.2
      - **SIGCHI Bulletin Vol.28 No.2, April 1996: Cognitive Architectures and HCI**
        http://www1.acm.org:82/sigs/sigchi/bulletin/1996.2/kirsch.html

        Preface:: *Issue Article Vol.28 No.2, April 1996 Article Issue Cognitive Architectures and HCI Susan S. Kirschenbaum, Wayne D. Gray, Richard M. Young Table of Contents Introduction Represented Architectures HCI Tasks References About the Authors The Cognitive Architectures and Human-Computer Interaction Workshop examined computational cognitive modeling approaches to human-computer interaction issues (HCI). The five major architectures and variations represented were briefly summarized. Part...*

        Persons: Adelson, B. || Authors, Addresses || Bauersfeld, P. || Bennett, J. || C.B. 344 || Dumais, S. || Erlbaum, Lawrence || Gray, Wayne D. || Kirschenbaum, Susan S. || Moran, Thomas P. || Olson, J. || P.G. (1994) ... (1995) || Pirolli, Peter || Young, Richard M

        Concepts: ACM CHI'92 Conference ... CHI'94 Conference ... Press || ACT-R & Soar ... model || ACT-R's conflict resolution mechanism ... rational analysis mechanism ... theoretical mechanisms || Analogical Reasoning || Ann Arbor || Appendix A || Applied Cognition program || Architecture ... bears || Architecture-defined boundaries || Architectures || Artificial Intelligence || Berlin || Boston || Boulder || Cambridge ... CB2 2EF || Chaucer Road || CHI conferences || Cognitive Architectures ... mechanisms ... Models ... phenomena ... Science ... task analysis ... theory || Colorado || Combinatorics problem || Comparing Architecture || Computational Model of Highly Interactive Task ... Model of Skilled Use of Graphical User Interface || Computer application ... Science || Computing Systems || Connectionist architectures || Construction-Integration theory || Cost-of-Knowledge Characteristic Function of information || Cricket Graph || D. W. & Krebs || Data search || Declarative memory chunks || Declarative/procedural memory distinction || Department of Electrical Engineering || Display-based human-computer interaction || Display-Based Systems || Division Newport || E. A. & Winterhalder || Engineering Psychologist || Environmental Contributions || EPIC Tech || Evolutionary Ecology || Experimental Psychology || Fairfax || Franzke (1994) || G. Lynch || George Mason University's Human Factors || Grain size || Graph-drawing program || HCI ... design problems ... issues ... phenomena ... problems ... task ... tasks || Human Behavior ... Factors || Human-computer interaction issues || Human-Computer Interaction journal ... Interaction Workshop || ICS Tech || Instantiations ... of tools || Institute of Cognitive Science || Interface Consistency || International Journal of Human-Computer Studies || Language text comprehension ... text processing system || LTM of queries || M. & Polson || Mason University || Memory element ... elements || Model Human Processor || MRC APU || NASA Ames Cognitive Modeling Workshop || Naval Undersea Warfare Center ... Undersea Warfare Center Division || New York || Newport || Perceptual motor performance problem || Ph.D. Dissertation || Princeton University Press || Proceedings of CHI 94 Conference ... of Sixteenth Annual Conference of Cognitive Science Society ... of Vienna Conference || Production firings ... per cycle ... system || Psychological Review || Psychology of Human-Computer Interaction || Raven || Research Scientist || Rieman (1993) || Skilled Interaction || Submarine || Table of Contents || Technical Report 94-02 || Teller machine || Two Cognitive Architectures || UK Medical Research Council's Applied Psychology Unit || University of Colorado ... of Michigan || Writing Desk

    - 1997.1
      - **SIGCHI Bulletin Vol.29 No.1, January 1997: AVI '96**

http://www1.acm.org:82/sigs/sigchi/bulletin/1997.1/avi96.html

Preface:: *Issue Article Vol.29 No.1, January 1997 Article Issue AVI '96 - An International Workshop Peter Pirolli The Third International Workshop on Advanced Visual Interfaces, AVI '96, was held May 27–29 in a converted monastery, in the town of Gubbio, in Umbria, Italy. Monks chose these golden hills of central Italy to build their cloisters, so that they could escape the plains of darker times. These hills exude savory foods and hearty wines that deeply satisfy the body so the mind ca...*

Persons: Alberti, Leon Battista || Costabile, Dr. Maria Francesca || Pirolli, Peter || Santucci, Dr. Giuseppe

Concepts: ACM Press ... SIGMultimedia || Advanced Visual Interfaces ... Visual Interfaces workshops || Animation techniques || Apennine mountains || Architecture || Associate Professor || AVI 96 ... 96 hosts ... 98 || Bari || Ben Scheiderman's call || Berkeley || Borgo San Sepolcro || Canada || Carnegie Mellon University || Chile || City states || Commune of Gubbio || Communication || Concrete applications || Coverage of works || Coyote Hill Road || Database research || Ecological perspective || Empricial methods || Esprit IDOMENEUS || Esprit's FADIVA || Europe || Festa dates || Generic design problem || Germany || Gran Sasso || Gubbio || Historic procession || Hotel Ai Cappuccini || Human-Computer Interaction || Information Foraging theory || International aspirations ... Workshop || Isabel Cruz of Tufts University || Italian AVI 96 hosts || Italy || Knolwedge content || Learning-to-learn strategies || Mathematics || Medieval dishes ... music ... times || Multimedia interfaces || Navigation || Olive trees || Palio della Balestra cross-bow competition || Palo Alto || PIE formalism || Portugal || Psychology || Renaissance || Roma La Sapienza || Rustic part of Italy || Saint Clare ... Francis || Sweden || Third International Workshop || Tilling of body || Torchlight || Umbrian countryside ... food ... towns || United Kingdom ... States || Universita di Roma || University of California ... of Toronto || Visual architectures ... databases ... interfaces ... Languages ... layout algorithms ... programming techniques || Web Forager || Workshop arrivals ... proceedings || Workspace || World Wide Web || Xerox Palo Alto Research Center ... PARC

cnidr { 1 }

k12 / wwwedu / 9503 /

- **search strategies, metaphors**
  http://k12.cnidr.org:90/wwwedu/9503/msg00155.html

dlib { 1 }

www / dlib / june96 / hearst /

- **Research in Support of Digital Libraries at Xerox PARC...**
  http://www.dlib.org/dlib/june96/hearst/06hearst.html

  Preface:: *Research in Support of Digital Libraries at Xerox PARC Part II: Paper and Digital Documents Marti Hearst, Gary Kopec, and Dan Brotsky Xerox PARC {hearst,kopec,brotsky}@parc.xerox.com...*

  Persons: Manero, Tony || Travolta, John

  Concepts: Algorithmic framework || Architecture || Brooklyn article ... disco || Butterfly || California dams || Cluster 4 || Communications media || Criminal Actions || D-Lib Magazine || Database record fields || Declarative specifications || Decoder generator || Degradation processes || Digital Documents ... images ... information ... library ... library context ... library project ... library projects ... library workshop ... materials ... world ... worlds || DIMSUM strategy || Distributional behavior of query terms || Document collections ... content ... groups ... Image Decoding Project ... Image Summarization activity ... model ... profiles ... recognition procedure ... recognition systems ... structure || Document-specific recognizers ... tags || Domain of closed-classed questions || Electronic behavior ... documents ... filing cabinet metaphor ... world || Encyclopedia || Englewood || Ethnographically-motivated design study || Financial institution || Fonts || Government articles || Information Access research ... retrieval ... retrieval systems || ISSN 1082-9873 || Java || Language analyses ... grammars || Lawyer's office || Libraries || Long Island || Murax Question

Answering || Navigation of retrieval results || New York City ... York City borough || Night Fever || OCR || Officers of Failed Financial Institutions || Online documents || Optical character recognition || Paper documents ... forms ... infrastructure ... server ... User Interface || PARC digital library technologies ... project ... work || Plane trip || Properties || Relational database entry || Retrieval of full document || Rocket engine development || Saturday Night Fever (1977) || Scatter/Gather Document || Semantic markup || SPARC20 workstation || SQL code || Stardom || Support library ... technologies || Term Set 1 ... Set 2 ... Set 3 ... Sets || Termsets || Text Database statistical content analysis engine || TileBars interface || Tipster collection || Topic 87 || Travolta article || TREC judges ... queries || TREC/TIPSTER collection || Typographic structure || UC Berkeley Digital Library project || User interface ... interface forms || Workspaces || Workstation interface || Xerox PARC